**The Dissertation Committee for Jacob Tinsley Whitt certifies that this is the approved version of the following dissertation:**

**DNA Target Site Recognition by the Ll.LtrB Group II Intron RNP**

**Committee:**

_____
**Alan M. Lambowitz, Supervisor**

_____
**Karen S. Browning**

_____
**Brent L. Iverson**

_____
**Tanya T. Paull**

_____
**Y. Whitney Yin**

# DNA Target Site Recognition by the Ll.LtrB Group II Intron RNP

**by**

**Jacob Tinsley Whitt, B.S.BioCh.**

**Dissertation**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin**

**August 2011**

# Dedication

To my partner, my parents, and my whole family for their

love and support through the years.

# Acknowledgements

Many people deserve my gratitude in helping to make this thesis possible.

First, I would like to thank my adviser, Dr. Alan M. Lambowitz, for the opportunity to work in his laboratory. His attention to detail and rigorous scientific reasoning have left a lasting impression on me.

I would also like to express my appreciation for the distinguished scientists on my graduate committee who provided their exptertise and time over the years, Dr. Karen S. Browning, Dr. Brent L. Iverson, Dr. Tanya T. Paull, and Dr. Whitney Yin.

The guidance of my mentor and friend Dr. James W. Noah early in my graduate career made everything that followed possible.

So many others in the Lambowitz Lab, too many to name individually, have been integral to my success and I feel fortunate to have been their colleague and friend.

I would also like to thank our collaborators, Dr. Wolfgang Frey, Dr. Soyeun Park, Dr. Jennifer S. Brodbelt, and Dr. Suncerae Smith, who provided instrumentation and instrumental expertise.

Finally, I would like to express my hearfelt gratitude to my entire family. My partner, Dr. Travis B. White, has been there for me every step of the way since I first set foot on the University of Texas campus. My mom, Wendy Whitt, has maintained a belief in me that has carried me through every self-doubt I have ever entertained. My dad, Tim Whitt, is among the hardest working and smartest men I have ever known and has always given me an example to strive for. My brother, Zach Whitt, has always been there for me, even as a kid. My grandparents, Ed and Mary Ann Derrich, instilled the appreciation for education and open-mindedness that led me to become a scientist. The influence of my grandparents, Delbert and Mary Lois Whitt, helped me to become the strong-willed and independent person I am. My sincerest thanks also go out to my entire family, my in-laws, and my friends for their love and support over the years.

# DNA Target Site Recognition by the Ll.LtrB Group II Intron RNP

Jacob Tinsley Whitt, Ph.D.

The University of Texas at Austin, 2011

Supervisor: Alan M. Lambowitz

Mobile group II introns are retroelements that site-specifically insert into DNA target sequences. The group II intron mobility pathway is mediated by a ribonucleoprotein particle (RNP) composed of excised intron RNA and an intron-encoded protein (IEP). The intron lariat inserts at a specific DNA target sequence and is then reverse transcribed by the IEP. Both the intron RNA and IEP are required for DNA target site recognition. I have identified the contact sites within the IEP responsible for recognition of two key positions in the DNA target, T+5 and T-23. IEP recognition of T+5 in the 3'-exon is required for endonuclease cleavage of the bottom-strand of the DNA target site, which generates a primer used for initiation of reverse transcription of the intron. The T+5 base is contacted by G498 in the LtrA DNA-binding domain and nearby residues, particularly K499, potentially bolster this interaction. Recognition of T-23 in the distal 5'-exon is required for initial recognition of the DNA target site by the RNP. The T533 side-chain contacts the T-23 base and the L534 side-chain may also contribute to recognition through hydrophobic interactions with the C5 methyl group. A

mutant, L534H, that switches target site specificity to T-23G has been characterized. In order for the RNP to make these and other contacts in the 5'- and 3'-exons simultaneously, the DNA must be bent. I have dissected the role of DNA bending in the intron mobility pathway and found that the DNA is bent at two progressively larger angles as the reaction proceeds. The predominant bend angle at earlier time points places the bottom-strand DNA cleavage site at the protein endonuclease active site. The predominant bend angle of later time points places the cleaved DNA site at the RT domain active site for initiation of reverse transcription of intron cDNA. Finally, in a practical application of group II intron mobility, I have used reprogrammed group II introns ("targetrons") to target two genes in *Bacillus subtilis* to demonstrate the suitability of targetron technology for gene targeting in the Gram-positive *Bacillus* genus.

# Table Of Contents

# List of Figures

# Chapter 1: Introduction

## 1.1 Mobile Group II introns

### 1.1.1 Introduction

Group II introns are fascinating ribozymes with a complex array of activities (Lambowitz and Zimmerly 2010). They self-catalyze two transesterification reactions for intron splicing to excise a branched intron lariat from an unprocessed transcript (Michel and Ferat 1995). They also catalyze the chemically identical reverse reaction to insert intron RNA into DNA. Sequence-specific reverse splicing of intron RNA into a double-stranded DNA target site provides the basis for retrohoming, the primary mechanism of group II intron mobility (Lambowitz and Zimmerly 2010).

Some group II introns are capable of self-catalyzing forward splicing in the presence of high $Mg^{2+}$ and/or salt concentrations, but splicing under physiological conditions requires the assistance of protein factors to facilitate intron folding (Lambowitz and Zimmerly 2010). Although proteins encoded within the genome of the host organisms can promote group II intron splicing (Jenkins et al. 1997; Perron et al. 1999; Till et al. 2001; Huang et al. 2005), intron-specific splicing factors ("maturases") encoded within the introns themselves can single-handedly facilitate splicing (Saldanha et al. 1999). The intron-encoded protein (IEP) promotes splicing by stabilizing the active conformation of the intron RNA during intron folding (Noah and Lambowitz 2003).

IEPs also play essential roles in group II intron mobility through DNA binding, reverse transcriptase, and site-specific DNA endonuclease activities (Zimmerly et al. 1995b; Guo et al. 1997). In retrohoming, the spliced intron RNA invades an intronless

allele and is reverse transcribed by the IEP, ultimately producing a new genomic copy of the intron (Zimmerly et al. 1995b; Cousineau et al. 1998). The intron RNA and IEP both contribute to retrohoming target site specificity (Guo et al. 1997). Mobile group II introns can also retrotranspose, invading ectopic DNA sequences, but less efficiently and at a lower frequency than retrohoming (Dickson et al. 2001; Ichiyanagi et al. 2002; Zhong and Lambowitz 2003).

The *Lactococcus lactis* LtrB (Ll.LtrB) intron and its cognate IEP, LtrA, provide a well-characterized model system for studying mobile group II introns. An efficient method for reconstituting Ll.LtrB RNPs *in vitro* using purified components has also been developed (Matsuura et al. 1997; Saldanha et al. 1999). The protein-assisted splicing and retrohoming pathways of the Ll.LtrB intron have been, and continue to be, the subject of exhaustive genetic and biochemical study. Targeting rules for the Ll.LtrB retrohoming DNA target site have been determined, and coupled with an understanding of the Ll.LtrB intron RNA contribution to DNA target site specificity, this has led to the development of an Ll.LtrB-based gene targeting system, the "targetron" (Mohr et al. 2000). Targetrons are capable of targeting virtually any gene in a variety of organisms (Karberg et al. 2001; Perutka et al. 2004; Yao et al. 2006; Yao and Lambowitz 2007; Mastroianni et al. 2008).

### 1.1.2 Group II intron structure

Group II introns fold into a conserved structure composed of six helical domains (DI-DVI) radiating from a central core with multiple long-range tertiary interactions between domains (Figure 1.1) (Michel and Ferat 1995; Qin and Pyle 1998; Lambowitz and Zimmerly 2010). DI provides the basis for active site assembly for both the splicing

and reverse splicing reactions. Together, DI and DV, which coordinates catalytically required divalent metal ions, form the minimal catalytic core of group II introns (Koch et al. 1992; Gordon and Piccirilli 2001). DII contributes to the assembly of the core active site during RNA folding through essential tertiary interactions. DIII, although not required for catalysis, is termed the "catalytic effector" and the reaction rate is reduced in its absence (Fedorova et al. 2003). The IEP open-reading-frame (ORF) is found in DIV along with a high-affinity binding site for the IEP. The binding site includes the ORF Shine-Dalgarno sequence, so that IEP binding regulates its own translation. The ORF itself is dispensable for intron catalysis (Wank et al. 1999; Singh et al. 2002). DVI contains a bulged adenosine residue, the branch-point adenosine, required for intron lariat formation (Michel and Ferat 1995).

Group II introns are divided into three subgroups, IIA, IIB, and IIC, based on specific structural variations including the number and arrangement of sequence elements in DI that are responsible for aligning the splice site at the intron active site (Lambowitz and Zimmerly 2010). The Ll.LtrB intron belongs to the IIA subgroup, which relies on base-pairing interactions from three sequence elements to align the exons in the active site for intron splicing. Exon binding sites 1 and 2 (EBS1 and EBS2, respectively) base pair with intron binding sites 1 and 2 (IBS1 and IBS2, respectively) in the 5'-exon, while δ base pairs with δ' in the 3'-exon (Figure 1.2). The same base-pairing interactions are also required for DNA target site recognition and reverse splicing in intron retrohoming (Lambowitz and Zimmerly 2010).

### 1.1.2 Group II intron splicing

Group II intron splicing proceeds via a reversible, two-step transesterification pathway (Figure 1.3), similar in mechanism to eukaryotic spliceosomal pre-mRNA splicing (Michel and Ferat 1995). The LtrA protein binds as a dimer to the partially folded Ll.LtrB intron RNA precursor at the high affinity binding site in DIV and makes secondary contacts in DI, DII, and DVI. LtrA binding stabilizes the active structure of the intron's catalytic core (Saldanha et al. 1999; Wank et al. 1999; Matsuura et al. 2001; Noah and Lambowitz 2003; Dai et al. 2008). In the fully-folded state, the 2'-OH of the branch-point adenosine is positioned to attack the phosphodiester bond of the 5'-intron-exon junction. This step generates a partially spliced intron lariat-3'-exon intermediate. The 3'-OH of the freed 5'-exon then attacks the phosphodiester bond of the intron lariat-3'-exon junction, simultaneously excising the intron lariat-IEP complex and ligating the 5'- and 3'-exons (Michel and Ferat 1995).

### 1.1.3 Group II intron retrohoming

Together, the excised intron lariat and IEP constitute the active ribonucleoprotein particle (RNP) required for intron mobility (Zimmerly et al. 1995a; Saldanha et al. 1999). Retrohoming predominantly follows a pathway in which the intron RNA reverse splices into a specific double-stranded DNA target site, followed by IEP endonuclease-dependent target-primed reverse transcription (TRPT). Cellular DNA repair mechanisms incorporate the cDNA-intron RNA into the host genome (Figure 1.4) (Lambowitz and Zimmerly 2010). The RNP initially binds DNA non-specifically and scans for the relatively long (~45 bp) specific target site. Upon recognition of the correct DNA

sequence features, reverse splicing is initiated by a conformational change within the RNP (Aizawa et al. 2003). Initial recognition by the RNP occurs in the 5'-exon and requires contacts between the LtrA protein and specific bases of the DNA, as well as phosphate backbone interactions (discussed in further detail in Section 1.3). These interactions between the 5'-exon and the LtrA protein precede localized melting of the double-stranded DNA, exposing IBS1, IBS2, and δ' in the top strand of the DNA target site to EBS1, EBS2, and δ in DI of the intron (Guo et al. 1997; Singh and Lambowitz 2001; Zhong and Lambowitz 2003). EBS-IBS and δ-δ' base pairing promotes reverse splicing by aligning the DNA insertion site with the intron RNA's active site.

After DNA target site recognition, the intron catalyzes the reversal of the two transesterification reactions used in intron splicing. In the first step of reverse splicing, the free 3'-OH of the intron attacks the phosphodiester bond of the insertion site in the DNA target top strand. The freed 3'-OH of the cleaved top strand DNA then attacks the 2'-5' phosphodiester of the intron lariat branch point (Lambowitz and Zimmerly 2010). As in intron splicing, the transesterification reactions are reversible, with the intron capable of inserting and excising itself from the DNA target. Downstream reactions mediated by the IEP provide the forward driving force for the mobility process (Aizawa et al. 2003).

During or after reverse splicing, LtrA makes the necessary contacts in the 3'-exon of the DNA target site required for endonuclease cleavage of the bottom strand of the DNA target and the TPRT reaction (Singh and Lambowitz 2001; Aizawa et al. 2003). LtrA cleaves the bottom strand nine nucleotides downstream from the Ll.LtrB intron insertion site and initiates cDNA synthesis of the intron RNA using the free 3'-end of the

cleaved bottom-strand DNA as the primer (Lambowitz and Zimmerly 2004). *In vivo*, host DNA recombination and repair processes which vary in different organisms complete retrohoming by incorporating the intron RNA-cDNA into the genome (Eskes et al. 1997; Smith et al. 2005; Zhuang et al. 2009).

### 1.2 LtrA, the Ll.LtrB intron-encoded protein

The LtrA protein consists of four domains: the reverse transcriptase (RT) domain, the maturase (X) domain, the DNA-binding (D) domain, and the endonuclease (En) domain (Figure 1.5). The N-terminal RT domain is composed of seven conserved motifs, RT1-RT7, with a large degree of homology to the fingers and palm regions of retroviral RTs. An additional upstream extension, RT0, is not found in long terminal repeat-containing (LTR-containing) retroviral RTs, but is conserved among the diverse family of non-LTR-retroelement RTs (which include the RTs of mobile group II introns, non-LTR-retrotransposons, bacterial retrons, and retroplasmids). RT0 is implicated in the binding of the RT to the RNA template for initiation of cDNA synthesis (Michel and Lang 1985; Xiong and Eickbush 1990; Chen and Lambowitz 1997; Malik et al. 1999; Bibillo and Eickbush 2002). Domain X is associated with activities related to RNA binding, namely, maturase activity to facilitate intron folding. It also contributes to RT activity. Although domain X shares little sequence homology to retroviral RTs, it contains secondary structural elements similar to the retroviral RT thumb and is located downstream of the RT domain in the position corresponding to the retroviral RT thumb and connection domains (Mohr et al. 1993; Blocker et al. 2005). Both the RT domain and domain X are required for protein-assisted intron splicing (Cui et al. 2004).

The C-terminal region of the IEP lacks homology or structural similarity to retroviral RTs. The DNA-binding domain contributes to reverse splicing and bottom-strand cleavage during intron mobility. Endonuclease cleavage of the bottom-strand requires the En domain. The En domain contains a $Mg^{2+}$-dependent nuclease fold homologous to the $Zn^{2+}$-dependent H-N-H endonuclease domain of colicin DNases (Singh and Lambowitz 2001; San Filippo and Lambowitz 2002). Mutations or deletions of conserved residues in the En domain that strongly inhibit En activity also inhibit RT activity to some degree, while RNA splicing and reverse splicing activities are largely unaffected. The relationship between En and RT activities may be explained by the need for proper positioning of the DNA primer by the En domain following bottom-cleavage so that the RT domain can initiate cDNA synthesis. However, the structure of the RT domain may also be sensitive to mutations in the En domain, analogous to the interdependent nature of the RT and RNase H activities of the HIV-1 RT (Ding et al. 1997), where mutations in one domain can have a large impact on the activity of the other (San Filippo and Lambowitz 2002; Zhong and Lambowitz 2003).

### 1.3 CRITICAL CONTACTS FOR LL.LTRB DNA TARGET SITE RECOGNITION

Group II intron RNPs recognize DNA target sequences by using both the IEP and base-pairing interactions of the intron RNA (Guo et al. 1997). Key nucleotide residues of the Ll.LtrB intron target site have been identified by mutational, DNA footprinting, modification-interference, and missing-base analyses (summarized in Figure 1.2) (Mohr et al. 2000; Singh and Lambowitz 2001). The IBS/δ' base-pairing region extends from positions -12 to +3, relative to the intron insertion site in the DNA top strand (Mohr et al.

2000). In both the distal 5'-exon and 3'-exon regions, the critical nucleotide residues recognized by IEPs differ even for closely related group II introns, suggesting that DNA target site specificity can evolve rapidly and may be readily changeable (Lambowitz and Zimmerly 2004).

### 1.3.1 IEP contacts in the distal 5'-exon

The most critical LtrA interactions in initial DNA target site recognition are found in the major groove of the distal 5'-exon extending from position -24 to -12, upstream of the IBS/δ' sequence (Figure 1.2). LtrA recognition of DNA features in this region of the DNA target site is required for both reverse splicing and endonuclease cleavage (Mohr et al. 2000; Singh and Lambowitz 2001).

Specifically, modification-interference using $KMnO_4$ identified T-23 and T-19, in the top strand, and T-20, in the bottom-strand, as significant pyrimidine contributors to 5'-exon target site recognition (Singh and Lambowitz 2001). $KMnO_4$ modification occurs at the C5-C6 double bond of the thymine base, which alters the major groove (McCarthy et al. 1990). Deoxyuridine substitution at T-23 strongly inhibits both reverse splicing and bottom-strand cleavage, while deoxyuridine substitution at the other two positions has no effect (Singh and Lambowitz 2001). Because deoxyuracil differs from thymine only in lacking the C5 methyl group, these findings suggest that initial LtrA protein recognition of the target site depends upon this distinguishing major groove feature of thymine at the -23 position.

Missing-base and dimethyl sulfate (DMS) methylation-interference experiments identified G-21 and A-20 as required contacts for both reverse splicing and bottom-strand

cleavage. Other purines in this region, G-17 and G-15 in the top strand and A-16 on the bottom strand, were not protected from nor sensitive to DMS methylation, despite effects from missing bases at these positions. They may be recognized specifically in the minor groove or may play a structural role in LtrA recognition. Phosphate ethylation-interference experiments indicate that the protein interacts non-specifically with the phosphate backbone along one face of the double helix in this same area (top strand -24 and -21, bottom-strand -15 to -19, and top strand -13 and -14). This region of the phosphodiester backbone is unusually susceptible to cleavage by 5-phenyl-1,10-phenanthroline in the absence of RNP binding and this may indicate an unusual backbone structure, which could contribute to recognition by the RNP. Overall, the results suggest that LtrA specifically recognizes T-23, G-21, and A-20 in the major groove of the top strand of the distal 5'-exon, bolstered by non-specific phosphate backbone interactions, early in the mobility reaction. These contacts are required for initiation of DNA melting to allow intron-DNA base pairing and/or stabilization of the DNA bottom strand as it becomes displaced by base pairing between the intron RNA and DNA top strand (Singh and Lambowitz 2001).

### 1.3.2 IEP contacts in the distal 3'-exon

Surprisingly few contacts between LtrA and nucleotide residues in the 3'-exon are highly sequence specific, with T+5 being the notable exception (Singh and Lambowitz 2001). Recognition of T+5 is not required for initial recognition of the target site, but it is stringently required for bottom-strand endonuclease cleavage later in the retrohoming pathway (Mohr et al. 2000). Nuclease accessibility and hypersensitivity at the T+5

position when the DNA target is bound by the RNP suggest that the base is unpaired prior to the appearance of substantial levels of bottom-strand cleavage products. Furthermore, $KMnO_4$ modification of T+5 or deoxyuridine substitution at T+5 both inhibit bottom-strand cleavage to similar degrees but have no affect on reverse splicing. Taken together, these results indicate that bottom-strand endonuclease cleavage requires IEP recognition of T+5 and the C5 methyl group of the thymine base contributes to the interaction between the IEP and T+5 (Singh and Lambowitz 2001).

LtrA interactions with other downstream positions in the 3'-exon top strand, T+11, T+12, and T+15, also contribute to bottom-strand endonuclease activity, but these are not strictly base specific and are not as critical as T+5 recognition. Interestingly, endonuclease cleavage of the bottom strand has no sequence specific requirement at the cleavage site itself (between the +9 and +10 positions) (Mohr et al. 2000; Singh and Lambowitz 2001).

### 1.3.3 The Role of the IEP C-terminus in DNA recognition

Biochemical and genetic studies using C-terminally truncated LtrA suggest that the C-terminal DNA-binding and En domains interact with the DNA target site at various stages of the retrohoming reaction (Singh and Lambowitz 2001; San Filippo and Lambowitz 2002). LtrA truncations in which both the DNA-binding and En domains are deleted abolish reverse splicing into double-stranded, but not single-stranded DNA targets. Deletion of only the En domain has no effect on reverse splicing into double- or single-stranded DNA targets but inhibits bottom-strand cleavage (San Filippo and Lambowitz 2002). Moreover, the En domain is not required for interactions with the

distal 5'-exon summarized in Section 1.3.1 and En domain point mutations that inhibit En activity do not prevent DNA target site recognition of the distal 5'-exon, DNA unwinding, or reverse splicing (Singh and Lambowitz 2001; Zhong and Lambowitz 2003). This pattern suggests that the DNA-binding domain is required for the local unwinding of distal 5'-exon DNA, and possibly for the stabilization of single-stranded DNA unwound by intron-DNA base pairing.

### 1.4 OVERVIEW OF DISSERTATION RESEARCH

This work addresses the mechanism used by the Ll.LtrB RNP to recognize and manipulate its specific DNA target sequence. In the first chapter, two sites found in the LtrA DNA-binding domain that interact with key DNA target sequence residues are identified. The peptide QGK (amino acid positions 497-499) contacts the T+5 base, while the peptide NTLENR (amino acid positions 532-537) contacts the T-23 base. To identify these contact sites, synthetic DNA target substrates with a site-specific photoactive nucleotide analog substitution, 4-thio-deoxythymine (4-S-dT), at either the -23 position, in the distal 5'- exon, or the +5 position, in the 3'-exon, were UV-cross-linked to the protein. The protein sites of the cross-links were then identified by electrospray-ionization ion-trap tandem mass spectrometry (ESI-ITMS/MS). Through mutational analysis of the identified recognition sites, these peptide contacts were confirmed and the nature of the interactions further characterized. Additionally, in light of the cross-linking results, a specificity switch LtrA mutant L534H was analyzed to better understand the nature of its altered specificity for the T-23G DNA target sites.

Chapter 2 examines the bending of the Ll.LtrB DNA target required to orient the DNA target successively into the intron active site for reverse splicing, the LtrA En domain active site for bottom-strand cleavage, and the LtrA RT domain active site for initiation of TPRT. I developed a model based on bending of the DNA target site as examined by atomic force microscopy (AFM). RNP-DNA complexes were immobilized on aminosilane-treated mica and imaged using tapping-mode AFM. This technique provides a snapshot of distinct molecular complexes at different points in the retrohoming reaction. By using a series of mutant RNPs and mutant or physically altered DNA substrates, the contributions of specific interactions to the bending of the DNA at different phases of the pathway were determined.

In a practical application of the work of the previous chapters on RNP target site interactions, Chapter 3 examines the suitability of the Ll.LtrB-based targetron as a gene targeting vector for the *Bacillus* genus. Functional targetrons were developed for two genes in *B. subtilis*, *ywpE* and *yhcS*. Homologs for both genes are found in the closely related, pathogenic species, *B. anthracis*. These genes encode sortase proteins responsible for attaching proteins to the cell wall of Gram-positive bacteria. Elimination of sortase genes in pathogenic Gram-positive bacteria, such as *B. anthracis*, attenuates virulence, making them attractive therapeutic targets (Weiner et al. 2010).

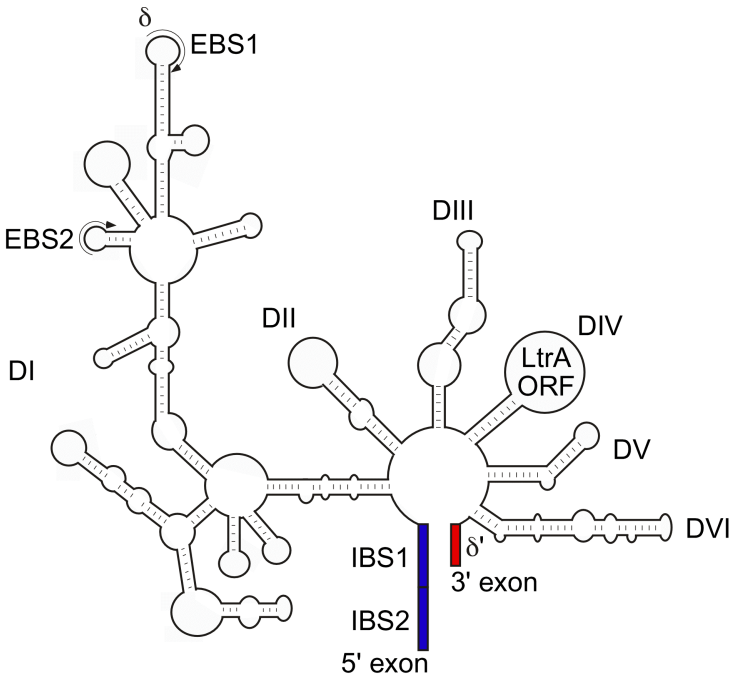


Figure 1.1: Secondary structure model of the Ll.LtrB group II intron

The six conserved helical domains of group II introns, labeled DI-DVI, emanate from a central hub. The model shown is specifically for the Ll.LtrB intron of the group IIA subclass, but the overall arrangement of the six domains is conserved among all three subclasses of group II introns. DI contains the sequence elements responsible for aligning the intron splice site, denoted EBS1, EBS2 and δ in group IIA introns, which base pair with IBS1 and IBS2 in the 5'-exon (blue) and δ' in the 3'-exon (red), respectively. DI and DV constitute the catalytic core of the intron, while DII and DIII contribute to intron folding and catalytic efficiency. The LtrA ORF and high affinity binding site are found in DIV. The bulged branch-point adenosine is located in DVI (Lambowitz and Zimmerly 2010).

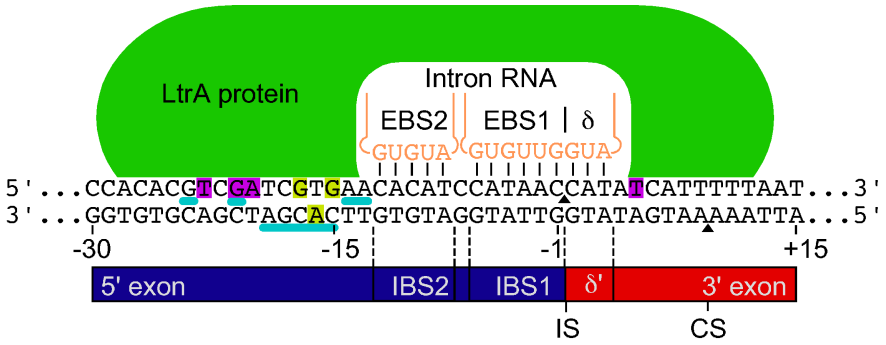


Figure 1.2: DNA target site interactions of the Ll.LtrB RNP

The 45-bp DNA target site (-30 to +15, relative to the intron insertion site) is shown in complex with the Ll.LtrB RNP which is composed of the LtrA protein (green) and the Ll.LtrB intron RNA (DI stem loop motifs containing EBS2 and EBS1/δ sequences shown in orange). The schematic below the DNA sequence denotes the IBS2 and IBS1 sequences in the 5'-exon (blue) and the δ' sequence in the 3'-exon (red), as well as the intron insertion site (IS) and endonuclease cleavage site (CS). The intron insertion site in the top strand of the DNA substrate and the endonuclease cleavage site in the bottom strand are also indicated by black arrows on the sequence. Nucleotide residues recognized by the LtrA protein through the major groove in the 5'-exon, T-23, G-21, and A-20, and T+5 in the 3'-exon, are highlighted in purple. Nucleotide residues that the protein contacts through the minor groove or non-base-specifically, G-17, G-15, and bottom-strand A-16, are highlighted in yellow. The LtrA protein interacts with the phosphate backbone on the 5' side of the positions underlined in cyan (Singh and Lambowitz 2001).

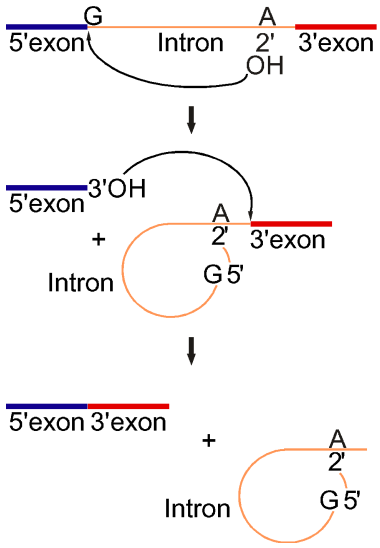


Figure 1.3: Group II intron splicing

Protein-assisted intron splicing and intron self-splicing follow the same two-step transesterification reaction pathway. Both steps are reversible. In the first transesterification reaction, the 2'-OH of the branch-point adenosine residue attacks the phosphate of the 5'-exon-intron junction (G residue) forming the partially spliced intron lariat-3'-exon intermediate. Splicing is completed by the second transesterification reaction with the attack of the intron lariat-3'-exon junction by the terminal 3'-OH of the 5'-exon, resulting in the ligation of the 5'- and 3'-exons and the release of the intron lariat (Lambowitz and Zimmerly 2010).

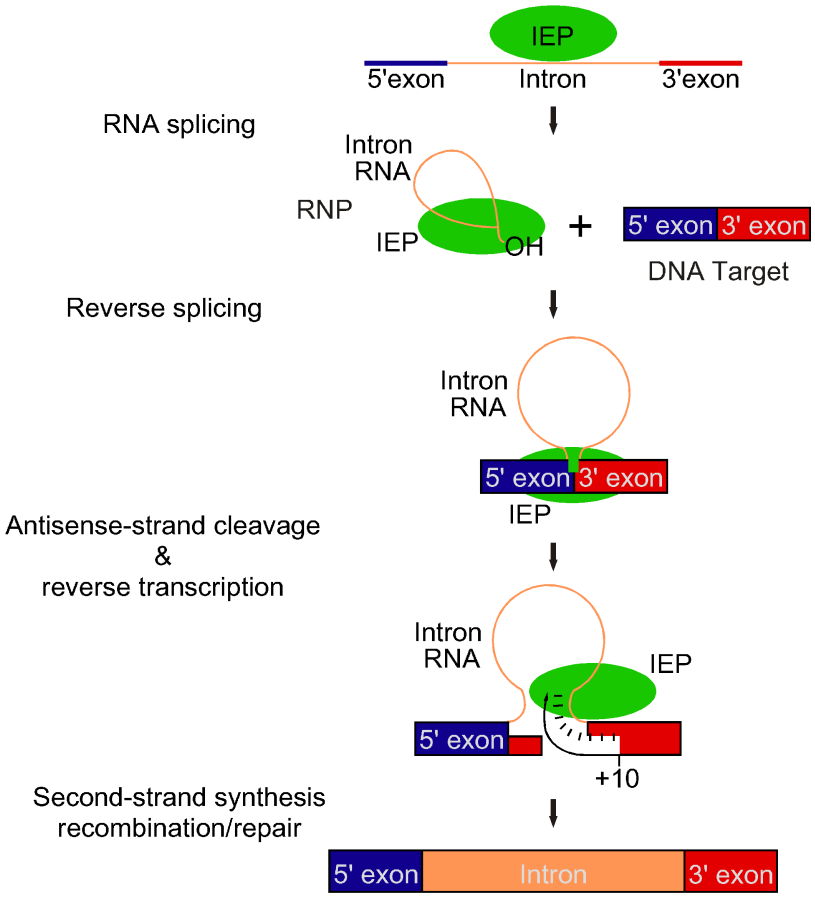


Figure 1.4: Group II intron retrohoming

The IEP binds the intron RNA in an unprocessed RNA transcript to promote splicing (see Figure 1.2). The IEP remains bound to the spliced intron lariat forming the retrohoming-competent RNP. The RNP initially recognizes the DNA target site through interactions between the IEP and the 5'-exon, which lead to localized DNA unwinding. Base pairing between the intron EBS/δ sequences and IBS/δ' DNA sequences align the intron active site for reverse splicing of the intron RNA into the top strand of the DNA target. The IEP cleaves the bottom strand between the +9 and +10 positions and uses the resulting 3'-OH to prime reverse transcription of intron cDNA (Lambowitz and Zimmerly 2010). Host DNA repair and recombination systems remove intron RNA and ligate the intron cDNA to genomic DNA to complete the retrohoming process (Eskes et al. 1997; Smith et al. 2005; Zhuang et al. 2009).

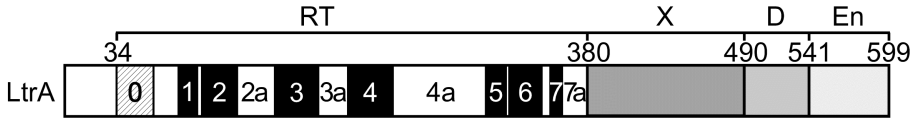


Figure 1.5: Domains of the LtrA protein

The Ll.LtrB IEP, denoted LtrA, is composed of four distinct domains, the reverse transcriptase (RT), maturase (X), DNA-binding (D), and endonuclease (En) domains, indicated above the protein schematic. Numbers above the schematic correspond to amino acid positions at domain boundaries. The highly conserved N-terminal RT domain is composed of seven conserved motifs (1-7) as well as an up-stream extension (0) which is conserved only among non-LTR-retroelement RTs. The RT domain and domain X bind the intron RNA, stabilizing its structure for splicing and reverse splicing. The C-terminal D and En domains interact with the DNA target site during intron mobility (Lambowitz and Zimmerly 2010).

## Chapter 2: Characterization of IEP-DNA Target Site Interactions

As shown in Figure 1.2, LtrA contacts T-23, G-21, and A-20 in the distal 5'-exon and T+5 in the 3'-exon of the DNA target site (Singh and Lambowitz 2001). To identify the contacts between specific amino acids of the LtrA protein and the DNA target site, I covalently cross-linked LtrA, in complex with the Ll.LtrB intron lariat, to DNA target sites containing a photoactive nucleotide analogue, 4-thio-deoxythymine (4-S-dT), at positions T-23 or T+5. Due to the longer activation wavelength (~365-nm) of 4-S-dT compared to native nucleotide residues, only cross-links between the specific DNA site and the LtrA protein were formed (Nikiforov and Connolly 1992; Wang and Biemann 1994). I then analyzed the resulting cross-link by electrospray-ionization collision-induced-dissociation ion-trap tandem mass spectrometry (ESI-CID-ITMS/MS) in the negative-ion mode. The identification of amino acids that interact with T-23 and T+5 provides the first opportunity to disrupt specific, known IEP-DNA target site contacts. I biochemically and genetically characterized the effects of disrupting these interactions by mutational analysis of the recognition sites.

Understanding the complex web of interactions required for DNA target site recognition by the IEP has potential gene targeting applications. The LtrA specificity switch mutant L534H was identified previously through unigenic evolution of segments of the LtrA DNA-binding domain, but the structural underpinnings of the specificity switch were unknown (San Filippo 2003). L534H supports intron mobility into T-23G mutant DNA target sites with higher efficiency than into the wild-type (T-23) DNA target site, which suggests that L534 is at or near the T-23 recognition site. Here, I show that

L534 is adjacent to a site of cross-linking between the protein (T533) and the -23 position of the DNA target site.

### 2.1 IDENTIFICATION OF KEY REGIONS OF THE DNA-BINDING DOMAIN

The DNA-binding domains of related group II IEPs, including LtrA, contain two structurally conserved features, an upstream basic cluster of amino acid residues and a downstream predicted α-helix (Figure 2.1) (San Filippo and Lambowitz 2002). The secondary structure prediction program Jpred 3 (Cuff and Barton 1999; Cole et al. 2008) predicts that the conserved α-helix of the wild-type LtrA DNA-binding domain spans N532-K539. This predicted length is somewhat shorter than the published predictions for the α-helix from previous versions of Jpred, which were from P524-L538 or A523-L538 (San Filippo and Lambowitz 2002; Blocker et al. 2005).

Key residues in the basic cluster and putative α-helix have previously been identified by screening for active LtrA variants from libraries of partially randomized ("doped") LtrA variants (San Filippo 2003). In the basic cluster, the most strongly conserved residue was R502, while K496, G498, R501, Y503, and, A505, were conserved but with greater numbers of substitutions. The most variable residues were Q497 and K499. In the predicted α-helical region, E535, R537, and L538 were invariant and most of the remaining amino acid residues were conserved. The two most variable residues were Q522 and N532. In both doped regions, most of the amino acid substitutions selected in active variants were conservative and/or similar to substitutions observed previously in unigenic evolution analysis: *e.g.*, T533S, N536D/I/T/K.

Notably, of the active LtrA mutants isolated from the α-helical region doped library, 55% contained a predicted α-helix spanning the positions predicted for the wild-type LtrA protein (positions 532-540). The core of the α-helix from positions 533 to 537 was predicted by the program for all active mutants in the sequenced library. Together, the above findings indicate that both the upstream cluster of basic amino acid residues and the predicted α-helical region are highly constrained in functional LtrA variants and provide additional insight into permissible amino acid substitutions in these regions.

## 2.2 THE T+5 CONTACT SITE

### 2.2.1 Identification of the T+5 cross-link site

To directly probe interactions between the LtrA protein and the 3'-exon of the DNA target site, I site-specifically photo-cross-linked LtrA to a modified DNA substrate. The DNA substrate used for cross-linking is a synthetic double-stranded DNA oligonucleotide corresponding to the 45-bp Ll.LtrB intron target sequence with 4-S-dT substituted at the +5 position. The 4-S-dT+5 DNA substrate was incubated with Ll.LtrB RNPs for 15 min at 37°C to allow reverse splicing and bottom-strand cleavage, and photo-cross-linked using 365-nm UV light (see Section 2.5.4 for experimental details). The covalently linked protein-DNA conjugates were then digested with a combination of DNase I and micrococcal nuclease, and precipitated with trichloroacetic acid/acetone to remove free nucleotides. The pellet was resolubilzed, denatured, and reduced with dithiothreitol (DTT) for trypsin digestion. The resulting mixture of tryptic-digest peptides was enriched for 4-S-dT-peptide cross-links using $TiO_2$ affinity chromatography. The enriched sample was then analyzed by mass spectrometry.

Figure 2.2 shows the full negative-ion mode ESI-ITMS of the enriched peptide-4-S-dT+5 cross-link mixture and the proposed cross-link structure. The peak at 634 m/z corresponds to the mass of the peptide QGK (330 Da), LtrA amino acid positions 497-499, plus the mass of the 4-S-dT monophosphate (4-S-dTMP) adduct (304 Da). This peptide sequence is found within the basic cluster in the LtrA DNA-binding domain (Figure 2.1) (San Filippo and Lambowitz 2002; Blocker et al. 2005).

Expected CID fragment ions are observed in the tandem MS (Figure 2.3) (Lenz et al. 2007). Because the $y_2$ backbone cleavage ion (508 m/z) includes the 4-S-dTMP adduct (Figure 2.3, inset), the cross-link is further localized to either G498 or K499. Cleavage occurs at the cross-link bond between the peptide and 4-S-dTMP adduct, releasing the unmodified peptide fragment (330 m/z). Fragmentation of the glycosidic bond produces the deoxyribose monophosphate ion (195 m/z) (Yousheng Hua 2000).

### 2.2.2 Cross-link and biochemical analysis of LtrA mutants at the T+5 contact site

To confirm that the cross-linked peptide is the contact site for the T+5 base, I constructed LtrA mutants in which alanine was substituted for the three amino acid residues of the QGK peptide, either individually or all together. Wild-type LtrA and the alanine-substituted mutants were then expressed, purified, and reconstituted into RNPs with Ll.LtrB RNA for biochemical analysis.

RNPs reconstituted with these alanine-substituted LtrA mutants were cross-linked to the 4-S-dT+5 DNA substrate containing a $^{32}$P-label at the +5 phosphate position (see Sections 2.5.3 and 2.5.4 for experimental details) to identify alanine substitutions that disrupt LtrA-T+5 interactions. After cross-linking, nucleic acids were DNase I digested

and the protein was analyzed by SDS-PAGE and autoradioagrapy (Figure 2.4). Cross-linking was observed by the transfer of the $^{32}P$ label to the protein. The cross-linking efficiency was unaffected by the Q497A mutation. However, the G498A and K499A mutations virtually abolished cross-linking, which suggests a disruption of the T+5 contact. The cross-linking efficiency of the QGK triple mutant was also significantly reduced. These findings agree with the MS data localizing the cross-link to either G498 or K499.

To ensure that the alanine-substitutions at these positions did not significantly alter tertiary interactions within the LtrA protein, these mutants were assayed for RT activity. The RT activity of purified wild-type and alanine-substituted LtrA was assayed by using the artificial template substrate poly(rA)/oligo$(dT)_{45}$ and quantifying polymerization of $^{32}P$-dTTP in the linear range (Figure 2.5). The results show that the RT activity of all four mutant proteins was as high or higher than that of wild-type LtrA, which indicates that these alanine substitutions did not disrupt global protein folding.

The reverse splicing and DNA endonuclease activities of these alanine-substituted LtrA mutants were determined in order to demonstrate the biochemical effects of disrupting the T+5 interaction (Figure 2.6). RNPs were incubated for 5 min at 37°C with an internally $^{32}P$-labeled, 129-bp DNA substrate containing the wild-type target site, and the products were separated in a denaturing 10% polyacrylamide gel (see Section 2.5.6 for experimental details). The Q497A, K499A and QGK substitutions inhibited both reverse splicing (92 ± 13%, 33 ± 5%, and 36 ± 16% of wild type, respectively) and bottom-strand cleavage (47 ± 16%, 8 ± 6%, and 8 ± 6% of wild type, respectively). The G498A substitution also inhibited bottom-strand cleavage (58 ± 4% of wild type), but this

was coupled with dramatically higher reverse splicing (242 ± 26% of wild type). This phenotype of decreased bottom-strand cleavage but increased reverse splicing has been observed previously with wild-type Ll.LtrB RNPs and DNA target sites having mutant nucleotide residues at T+5 (Mohr et al. 2000; San Filippo and Lambowitz 2002) and, similarly, for the yeast mitochondrial group II intron aI2 when single nucleotides from the 3'-exon of its DNA target site (+1 to +4) were deleted (Guo et al. 1997). Together with the analytical cross-linking data, this phenotype strongly suggests that the G498A mutation specifically prevents normal protein interaction with T+5.

### 2.2.3 Genetic analysis of LtrA mutants at the T+5 contact site

To further characterize the effects of mutating LtrA positions 497 through 499, the ability of these alanine-substituted LtrA mutants to support *in vivo* intron mobility was assessed using the *E. coli* two-plasmid assay. The *E. coli* two-plasmid assay, diagrammed in Figure 2.7, utilizes the $Cam^R$ intron-donor plasmid pACD2X and an $Amp^R$ recipient plasmid pBRR-ltrB to measure the efficiency of intron mobility. The intron/IEP donor plasmid uses a lactose-inducible T7 promoter ($P_{T7lac}$) to express an Ll.LtrB intron RNA with flanking 5'- and 3'-exons. A T7 promoter ($P_{T7}$) is inserted in DIV of the intron near the 3'-end, which ultimately provides a marker for intron insertion into the recipient plasmid (described below). The LtrA protein is also expressed from pACD2X, downstream of the Ll.LtrB 3'-exon. The recipient plasmid, pBRR-ltrB, harbors a promoterless $tet^R$ gene downstream of the Ll.LtrB intron target site (ligated 5'- and 3'-exons). Induction of the donor plasmid leads to expression of an Ll.LtrB precursor RNA, which is spliced by the LtrA protein to produce active RNPs. RNP-mediated intron

integration into the DNA target site of the recipient plasmid delivers the $P_{T7}$ promoter and activates *tet*$^R$ expression. The mobility efficiency is quantified as the ratio of Amp$^R$ + Tet$^R$ colonies to Amp$^R$ colonies.

The mobility efficiency of the Ll.LtrB intron was assayed using recipient plasmids containing a target site orientated so that either the nascent lagging (pBRR3A-ltrB, LAG) or leading (pBRR3B-ltrB, LEAD) strand of replication could potentially be used as a primer for reverse transcription of the intron cDNA in the absence of endonuclease cleavage (Figure 2.8). A mobility bias towards the LEAD recipient has been used to discern LtrA mutations that specifically inhibit endonuclease cleavage (Zhong and Lambowitz 2003). As a control, the YRT LtrA mutant was also assayed in parallel. The previously characterized YRT LtrA mutant, which has three point mutations (Y529A, R531A, and T533A), is endonuclease deficient with slightly higher than wild-type RT activity (117%) and reduced reverse splicing activity (59%) (San Filippo and Lambowitz 2002). The mobility frequency of YRT LtrA is greater with recipient vectors that allow the nascent leading strand from replication to prime reverse transcription of intron cDNA, indicating inhibition of bottom-strand cleavage activity which prevents En-mediated TPRT (Zhong and Lambowitz 2003).

The degree of *in vivo* intron mobility inhibition by the Q497A, G498A, K499A, and QGK mutations parallels the *in vitro* biochemical activities found for each mutant as described in the previous section (Figure 2.9). The intron mobility efficiency using the G498A LtrA mutant (LEAD 43 ± 2%; LAG 37 ± 4%) was approximately wild type (LEAD 43 ± 1%; LAG 44 ± 2%), reflecting G498A LtrA's relatively high *in vitro* reverse splicing activity with only slightly diminished endonuclease cleavage activity. Q497A

LtrA supported slightly lower mobility than wild-type LtrA (LEAD 32 ± 2%; LAG 29 ± 1%), also in line with the *in vitro* biochemistry data. The mobility efficiencies using K499A LtrA (LEAD 23 ± 1%; LAG 21 ± 1 %) and QGK LtrA (LEAD 31 ± 1 %; LAG 24 ± 1%) were also significantly lower than wild-type LtrA, mirroring the trend of the *in vitro* biochemical data. In agreement with previous results (San Filippo and Lambowitz 2002; Zhong and Lambowitz 2003), the mobility efficiency using YRT LtrA, which is much lower than that of wild-type LtrA with either recipient plasmid, was strongly biased towards the LEAD recipient (LEAD 0.7 ± 0.09%; LAG 0.3 ± 0.02).

Although none of the alanine substitutions at LtrA positions 497 through 499 produced a pronounced LEAD replication orientation bias in intron mobility, intron mobility with these LtrA mutants compares favorably to the *in vitro* biochemistry results. These data indicate that the inhibition of bottom-strand cleavage by the Q497A, K499A, and QGK mutations may be due to effects on protein structure that inhibit both reverse splicing and DNA endonuclease activity. The data also indicate that the increase in reverse splicing activity of G498A LtrA is enough to overcome its slightly reduced endonuclease cleavage activity to give a wild-type level of *in vivo* mobility. As implied by the *in vitro* biochemical data, the Q497A, G498A, K499A and QGK mutations only partially disrupt the T+5 interaction so that the endonuclease activity of the these mutants is not inhibited enough to observe the pronounced LEAD replication orientation bias in intron mobility. In contrast to the G498A LtrA mutant, YRT LtrA mutant phenotype does not completely match the expected phenotype for an LtrA mutant that does not recognize T+5 (San Filippo and Lambowitz 2002; Zhong and Lambowitz 2003). Lack of T+5 recognition is expected to both stimulate reverse splicing and inhibit bottom-strand

cleavage, as demonstrated when intron mobility with wild-type LtrA was assayed using T+5 mutant target sites (Mohr et al. 2000).

## 2.3 THE T-23 CONTACT SITE

### 2.3.1 Identification of the T-23 contact site

To identify the LtrA site of interaction with T-23 in the DNA target site, the protein was photo-cross-linked to a DNA substrate containing the 4-S-dT substitution at the -23 position, in the same manner described for T+5. Figure 2.10 shows the full negative-ion mode LC-ESI-ITMS of the cross-link-enriched tryptic-digest of the 4-S-dT-23-peptide cross-link. The spectrum shows a doubly-charged peak at 522 m/z (singly-charged m/z = 1045) corresponding to the peptide NTLENR, LtrA amino acid positions 532-537, with the 4-S-dT adduct mass addition (Figure 2.10, inset). This sequence is found in the conserved putative α-helix of the LtrA DNA-binding domain (Figure 2.1) (Blocker et al. 2005). This cross-linked peptide required further separation from non-cross-linked peptide background and other contaminants by in-line RP-HPLC following $TiO_2$-affinity enrichment for MS detection, which enhanced the relative signal of the cross-linked species.

A species resulting from γ fragmentation of the backbone, which is dependent on the asparagine side chain (N536), is observed in the tandem mass spectrum (Figure 2.11) (Bowie et al. 2002). Phosphate loss from fragments containing the 4-S-dT adduct also occurs (Lenz et al. 2007), but the low-mass limit of this spectrum does not allow detection of the deoxyribose monophosphate species itself, which is the glycosidic bond cleavage fragment as seen in Figure 2.3 for the T+5 cross-link. Fragments resulting from

signature neutral losses from the side chains of asparagine, threonine, leucine, glutamic acid and arginine indicate the presence of these amino acids, but not their sequence order (Figure 2.11) (Waugh et al. 1990; Waugh et al. 1991; Waugh et al. 1993; Bowie et al. 2002). The mass of the fragment resulting from the threonine side-chain loss indicates that the cross-link is formed between the side-chain hydroxyl of T533 and the 4-S-dT residue (Waugh et al. 1990).

### 2.3.2 Biochemical activity of LtrA mutants at the T-23 contact site

To confirm the T-23 contact, LtrA variants containing single alanine substitutions for each amino acid residue from N532 through R537 in the cross-linked peptide were constructed as described above for the Q497-K499 peptide. With the exception of L534A LtrA, the single alanine substitutions at positions N532 through R538 caused the protein to express poorly. The RT activity of these purified alanine-substituted mutants, again with the exception of L534A LtrA, assayed with the artificial template-primer substrate poly(rA)/oligo$(dT)_{45}$, as described above, was also inhibited (Figure 2.12). Not surprisingly, all of the RT-inactive mutants were inactive in the reverse splicing and endonuclease cleavage assays as well (Figure 2.13). Because these alanine substitutions in the DNA-binding domain of LtrA lead to inhibition of RT activity, which is primarily dependent on the RT and X domains, they presumably disrupt global protein structure. Therefore, it cannot be determined from these biochemical data whether the N532A, T533A, E535A, N536A and R537A mutations specifically disrupt recognition of T-23 in the DNA target site. The L534A mutation may also affect global LtrA structure but the disruption of the tertiary structure of the L534A LtrA mutant is apparently less severe

compared to that of the neighboring alanine-substitution mutants based on their relative RT activities (Figure 2.12).

Interestingly, the partial reverse splicing activity of L534A LtrA was significantly decreased (27 ± 5% of wild type), but full reverse splicing undiminished (99 ± 6% of wild type) (Figure 2.13). Additionally, endonuclease activity was severely inhibited by the L534A mutation (11 ± 1% of wild type), which is not expected for a mutant that supports reverse splicing at wild-type levels and makes normal contacts in the 3'-exon. The L534A mutation may affect binding of the DNA target site in a way that alters the kinetics of reverse splicing and later steps of the mobility reaction. The effects on reverse splicing and endonuclease cleavage could be either in addition to, or because of, changes in the recognition of T-23. Initial target site recognition, which is dependent on the T-23 contact, leads directly to the first step of reverse splicing. Because the amount of the partially reverse spliced intermediate is reduced compared to wild-type LtrA, the accumulation of the fully reverse spliced intermediate suggests that the reversibility of the first transesterification reaction of reverse splicing is affected by the L534A mutation. The slightly altered tertiary structure of L534A LtrA may indirectly disrupt protein contacts with the 3'-exon by positioning the bound DNA target site in way that prevents other sites in the protein from interacting with T+5 or the bottom-strand cleavage site.

### 2.3.3 Genetic analysis of LtrA mutants at the T-23 recognition site

Although L534A was the only active mutant from the conserved α-helical region that expressed and purified normally, results from LtrA unigenic evolution experiments implied that these mutants may be active in the *in vivo* mobility assay when expressed

simultaneously with Ll.LtrB intron RNA (San Filippo 2003). Mutations that affect protein recognition of T-23 in the DNA target site are not expected to lead to any replication orientation bias in mobility frequency. However, the YRT LtrA mutant (discussed in detail in Section 2.2.3) partially overlaps with the conserved α-helical region and does show a replication orientation bias in the mobility assay. Therefore, the N532-R537 alanine-substituted LtrA mutants were also tested for the bias in mobility.

The single alanine substitutions at positions N532 through R537 strongly inhibit intron mobility (Figure 2.14). The intron mobility efficiency with the N532A LtrA mutant was particularly low (LEAD 0.52 ± 0.07%, LAG 0.08 ± 0.03%). Alanine substitutions at the other positions inhibited intron mobility less severely but were still uniformly low (T533A: LEAD 12 ± 1%, LAG 22 ± 1%; L534A: LEAD 10 ± 0.2%, LAG 9 ± 1%; E535A: LEAD 12 ± 0.3%, LAG 12 ± 1%; N536A: LEAD 17 ± 1%, LAG 23 ± 1%; R537A: LEAD 16 ± 6%, LAG 6 ± 1%). Taken together with the *in vitro* biochemical data, these findings suggest that single alanine substitutions at positions N532, T533, E535, N536, and R537 disrupt protein tertiary structure but this effect can be at least partially overcome by the presence of the Ll.LtrB intron RNA as the protein is translated and folds *in vivo*. Some fraction of L534A LtrA is capable of folding properly without co-expression of the Ll.LtrB intron RNA based on its combination of higher *in vitro* biochemical activity with similar *in vivo* mobility efficiency as the other N532-R537 alanine-substituted mutants.

Interestingly, the N532A and R537A LtrA mutants produced a leading strand mobility bias. Because such a mobility bias reflects a specific inhibition of bottom-strand cleavage but not reverse splicing, this phenotype is unexpected for mutations near the site

of T-23 recognition. This may be caused by the disruption of the protein's normal tertiary interactions, which were also implicated in the *in vitro* biochemical analysis. These mutations could also directly alter protein binding to the DNA substrate to allow reverse splicing but inhibit bottom-strand cleavage by mispositioning the 3'-exon relative to the En domain. A scenario where both disrupted protein tertiary structure and direct effects on DNA substrate binding contribute to the observed phenotype is also possible.

### 2.3.4 T-23 to G-23 specificity switch

The same partially-randomized libraries used for the mutagenic evolution experiments (see Section 2.3.1) were tested against DNA target sites with mutant -23 or -21 positions to identify potential specificity switch LtrA variants. This screening identified two variants with higher mobility efficiencies against the T-23G target site compared to the wild-type target site. Both LtrA variants contained the mutation L534H (one also contained an additional mutation, N532T), and two other variants with mobility efficiencies similar to wild-type LtrA with the T-23G target site contained the mutation L534H along with other mutations (San Filippo 2003).

Based on these findings, I constructed an LtrA variant with the point mutation L534H, and carried out genetic assays comparing the mobility efficiencies of wild-type LtrA and L534H LtrA with -23 position mutant DNA target sites (Figure 2.15). In agreement with previous results, the wild-type LtrA protein showed a strong preference for the wild-type DNA target site containing T-23 and decreased mobility frequencies with the three mutant target sites (30-50% relative to wild-type target site) (Zhong et al. 2003). By contrast, L534H LtrA showed a marked specificity switch, with a preference

for T-23G (49 ± 3% relative to wild-type LtrA with wild-type target site) compared to the wild-type or other mutant target sites (20-30% relative to wild-type LtrA with wild-type target site).

## 2.4 DISCUSSION

### 2.4.1 Functionally important regions of the DNA-binding domain

Multiple sequence alignments of the LtrA C-terminal domains with those of related group II IEPs show little sequence conservation within the DNA-binding domain (San Filippo and Lambowitz 2002; Blocker et al. 2005). However, two functionally important regions were identified and confirmed by unigenic evolution analysis, a high-throughput genetic method in which functional variants are isolated from a library of mutant proteins with random PCR-induced mutations (San Filippo and Lambowitz 2002). One of these regions is near the beginning of the DNA-binding domain and contains a small cluster of basic amino acid residues. The other is near the N-terminus of the domain and contains a predicted α-helix, which was conserved in all functional variants (Figure 2.1). Despite minimal sequence homology, both of these regions have been identified in the DNA-binding domains of related group II intron IEPs.

### 2.4.2 T+5 recognition

My MS results map a T+5 contact site to the peptide QGK (positions 497-499) in the LtrA DNA-binding domain. The collision induced fragmentation pattern indicates that the cross-linked amino acid residue is G498 or K499. Biochemical experiments confirmed that alanine substitutions at G498 or K499 virtually abolish cross-linking to 4-

S-dT+5, and both mutations inhibit DNA endonuclease activity. The Q497A mutation has little effect overall. Moreover, the G498A substitution creates the same distinctive phenotypic pattern as a T+5 mutation, inhibiting bottom-strand cleavage but stimulating reverse splicing (Mohr et al. 2000). This finding strongly suggests that G498 is involved in recognition of T+5, while K499 may strengthen the contact.

The simplicity of glycine provides for limited routes of interaction with the thymine base. In some cases, the hydrogen of the peptide nitrogen can form a hydrogen bond with the thymine O2, though this provides no strict sequence specificity because any base provides an H bond acceptor at roughly the same position (Treiber et al. 2010). A more likely scenario for the glycine-T+5 base interaction is through the α-carbon and side-chain hydrogen, which can pack closely with the thymine C5 methyl group through hydrophobic interactions, as seen in the helix-turn-helix (HTH) motif of the lambda repressor bound to its operator sequence (Jordan and Pabo 1988). This potential mode of interaction is supported by deoxyuracil substitution and $KMnO_4$ modification interference data that point to features of T+5 on the major groove side of the base, particularly the C5 methyl group, as important determinants for recognition by LtrA.

Curiously, while G498 is strongly selected in active LtrA variants, the two flanking residues Q497 and K499 are the most variable in the basic cluster region (San Filippo 2003). The function of the neighboring residues in the basic cluster is unclear. Previous work analyzed the effect of mutations at R501, R502, and Y503 (San Filippo and Lambowitz 2002). This work showed that mutating R501, R502, and Y503 all to alanines (the RRY mutant) inhibits both reverse splicing and bottom-strand cleavage activity. The R501A and R502A single mutations also inhibit both reverse splicing and

bottom-strand cleavage. However, the reverse splicing and bottom-strand cleavage of Y503A LtrA are at near wild-type levels. Of the four, only R502A LtrA retains near wild-type RT activity (83%). This pattern suggests that these amino acids may contribute to DNA binding through some partially redundant interactions which help coordinate the LtrA En and RT domains for bottom-strand cleavage and initiation of TPRT, respectively, or they may be involved in the LtrA conformational changes required for transfering the DNA cleavage site to the RT active site to initiate TPRT following endonuclease cleavage as discussed in Section 3.3 (San Filippo and Lambowitz 2002; Noah et al. 2006). The T+5 contact and basic cluster may help anchor the En domain in the correct position to cleave between positions +9 and +10.

### 2.4.3 T-23 recognition

Protein recognition of T-23 is thought to occur early in the intron mobility reaction and is required for reverse splicing. Deoxyuracil substitution and $KMnO_4$ modification data also indicate that LtrA contacts T-23 in the major groove through the C5 methyl group (Singh and Lambowitz 2001). Here, I have identified a predicted α-helix, spanning the peptide NTLENR in the LtrA DNA-binding domain, responsible for making the T-23 contact. Protein α-helices often make base-specific contacts through the DNA major groove (Church et al. 1977; Luscombe and Thornton 2002). The geometry of the major groove accommodates α-helices in multiple orientations which allows for a wide variety of potential interactions (Pabo and Sauer 1992). This putative α-helix (N532A-K539) is long enough, at approximately 12 Å, to make contact with additional nearby positions in the distal 5'-exon of the DNA target site found to be important for sequence recognition,

such as G-21 or A-20, or with sites on the phosphate backbone of either the top or bottom strand.

Using the cross-linking, biochemical and genetic data presented here, a model of the interactions between this α-helix and the DNA target site has been created by hand docking an idealized α-helix encompassing this peptide onto the B-form DNA substrate (Figure 2.16). Based on the nature of the 4-S-dT-23-peptide cross-link, the hydrogen of the T533 side-chain hydroxyl is in position to interact with O4 carbonyl of the thymine base. The mutational analysis and specificity switch genetic data imply that the neighboring residue, L534, may also be involved in recognition of T-23. Multiple orientations of the α-helix could allow for the T533-T-23 interaction, but only one potential orientation also places the hydrophobic L534 side-chain near the thymine C5 methyl group. In this same orientation R537 and K539 are positioned to contact the phosphate backbone at positions -24 in the top strand and -19 in the bottom strand, respectively, which have both been indicated as sites of contact by ethylation interference data (Singh and Lambowitz 2001). Additionally, the E535 side-chain is positioned to potentially interact with either G-21 or A-20, depending on the specific side-chain rotamer as it exists in the actual α-helical structure. Recognition of guanosine or adenosine bases by glutamic acid has been shown to occur through direct hydrogen bonding or indirect water-mediated bonding (Luscombe et al. 2001), either of which is possible in this model.

This model also provides a potential explanation for the L534H LtrA specificity switch to T-23G (Figure 2.17). Substitution of L534 with histidine sterically prevents the same T533 interaction with a thymine base at the -23 DNA position, even if the α-helix is

reoriented to account for the larger histidine side-chain. Furthermore, slightly rotating the α-helix brings the histidine side chain into close enough proximity to the guanine base at the -23 position to interact through hydrogen bonding, the typical mode of interaction for histidine and guanine (Garvie and Wolberger 2001; Luscombe et al. 2001). The reoriented α-helix is also still positioned to allow the other potential contacts described above, but with less ideal geometries. This may explain the generally reduced mobility efficiency of the L534H mutant compared to wild-type LtrA.

### 2.4.4 Analysis of DNA-binding domain site LtrA mutants

Previously, the substitution of Y529, R531, and T533 all with alanines (the YRT mutant) was found to inhibit reverse splicing and even more strongly inhibit bottom-strand cleavage, with Y529A the most critical of the three mutations. R531A LtrA retains high RT activity with only somewhat reduced reverse splicing and bottom-strand cleavage activity. T533A LtrA had wild-type levels of both activities in the previous study, however, here I find that T533A LtrA is virtually inactive in RT, reverse splicing and bottom-strand cleavage assays. This contradiction may be due to differences in the purification methods used to isolate the recombinant LtrA proteins. The current method includes a polyethylenimine (PEI) precipitation to remove cellular nucleic acid non-specifically bound to LtrA that would be carried through the purification process. Previously, PEI precipitation was not used to remove nucleic acid.

Unlike G498A LtrA, YRT LtrA did not show stimulation of reverse splicing activity in concert with decreased bottom-strand cleavage activity. The conclusion was that the pattern of inhibition for YRT LtrA was consistent with a disruption of interactions that

contribute to both reverse splicing and bottom-strand cleavage. The stronger inhibition of bottom-strand cleavage by these mutations could reflect more severe disruption of 3'-exon interactions that are required specifically for bottom-strand cleavage or disruption of interactions with the 5'-exon which result in an altered orientation of the En domain that prevents bottom-strand cleavage.

This idea is supported by the reverse splicing and bottom-strand cleavage data for alanine substitution at the nearby L534 position. Partial reverse splicing and bottom-strand cleavage are both inhibited by the L534A mutation. However, the apparent level of full reverse splicing, the reaction step between partial reverse splicing and bottom-strand cleavage, is at wild-type levels. This finding raises the possibility that the positioning of the 5'-exon upon initial recognition is affected by the mutation and this altered binding affects the equilibrium between the partial and full reverse splicing intermediates and downstream endonuclease cleavage. Additionally, the mobility efficiencies of the N532A and R537A LtrA mutants display a replication orientation bias towards the LEAD recipient, indicating a specific inhibition of bottom-strand cleavage. This unexpected phenotype for mutants so close to the site of T-23 contact at residue T533, underscores the role that amino acid residues near the contact sites in the DNA-binding domain play in orienting other specific protein domains correctly with respect to the DNA substrate for all phases of the retrohoming reaction pathway.

Although I demonstrate that G498/K499 and T533 interact directly with the DNA target site, other nearby amino acid residues may be involved in direct or indirect interactions with the En or RT domains. DNA-binding domain mutations are known to affect RT activity of group II intron IEPs (San Filippo and Lambowitz 2002). YRT LtrA

has slightly higher than wild-type RT activity, as does the yeast aI2 IEP mutant P714T, which contains a similarly placed point mutation in the analogous predicted α-helix (Zimmerly et al. 1995b). The RT activity of Y529A LtrA and R531A LtrA is stimulated to an even greater degree (169% and 157% of wild type, respectively) (San Filippo and Lambowitz 2002). Conversely, deletion of the En domain essentially abolishes RT activity, while En domain point mutations at conserved amino acid positions are, to varying degrees, inhibitory (San Filippo and Lambowitz 2002). These effects could be due to the disruption or strengthening of direct interactions between the domains of the IEP or caused by altered positioning of the DNA during the mobility reaction and initiation of reverse transcription.

### 2.4.5 Specificity switch LtrA mutants

The identification of specific amino acid residues in the LtrA protein that interact with important features of the DNA target site during retrohoming may have practical applications in gene targeting. Although LtrA recognition of specific DNA target site positions is flexible enough to identify multiple target sites in most genes, the range of target sites and the mobility efficiency at some potential target sites is certainly limited by these fixed protein interactions (Perutka et al. 2004).

Here, I have identified the site of interaction with T-23 in the DNA target site as LtrA amino acid residue T533. Previously, the L534H mutation at the neighboring residue was identified as causing a specificity switch to T-23G. However, as demonstrated here, the degree of specificity for G-23 by L534H LtrA is not as great as that of wild-type LtrA for the wild-type (T-23) DNA target (Figure 2.15), which suggests

potential room for improvement. Mutations that compensate for the altered orientation of the conserved α-helix with the L534H mutation may make it possible to increase both the specificity for T-23G and the overall retrohoming efficiency.

No similar LtrA specificity switch mutant for T+5 has been identified thus far. Now that G498 has been identified as the T+5 LtrA interaction site, work has begun in this direction. Even if the selectivity for a specific nucleotide residue at -23 or +5 of an LtrA mutant is not such to make it a true specificity switch, simply relaxing the targeting rules at these positions may also be useful. More flexible recognition at these DNA positions by the IEP simplifies the task of identifying potential target sites and may increase the number of possible target sites within a targeted gene.

## 2.5 METHODS

### 2.5.1 Recombinant plasmids and mutant LtrA constructs

pACD2, the intron donor plasmid used for genetic assays, contains a 940–nt Ll.LtrB–ΔORF intron with a T7 promoter inserted in intron domain IV (Guo et al. 2000; Karberg et al. 2001). The intron and flanking exons are cloned downstream of an IPTG–inducible T7lac promoter in a $Cam^R$ pACYC184–derivative, with the LtrA protein expressed from a position downstream of the 3'-exon. pACD2X is a derivative of pACD2 that contains an XhoI site inserted 8-bp downstream of the LtrA ORF termination codon to facilitate introduction of mutations (San Filippo and Lambowitz 2002).

pBRR3–ltrB, the recipient plasmid used for genetic assays, contains a minimal Ll.LtrB intron target site (ligated *ltrB* gene E1–E2 sequence extending from positions –30 to +15 from the intron–insertion site) cloned upstream of a promoterless $tet^R$ gene in an

$Amp^R$ pBR322–derivative (Guo et al. 2000; Karberg et al. 2001). Recipient plasmids containing mutant target sites were constructed by synthesizing complementary DNA oligonucleotides that correspond to the modified DNA target sequence flanked by AatII and EcoRI sites. The annealed DNA oligonucleotides were digested with AatII plus EcoRI and cloned between the corresponding sites of similarly digested pBRR3–ltrB, which had been treated with calf intestinal alkaline phosphatase (CI–AP) (Invitrogen, Carlsbad, CA).

Alanine-substituted LtrA mutants for *in vitro* use were created by generating double-stranded DNA inserts containing GCA codon(s) substituted in-frame into the LtrA ORF by two-step, overlapping PCR using pImp-1P as template (Saldanha et al. 1999). The inserts were flanked by XmaI and MscI restriction sites. Full-length inserts were digested with XmaI and MscI and purified in a 1% agarose gel. pImp-1P was digested with XmaI and MscI and the fully digested vector backbone was agarose gel-purified. Full-length, digested inserts were ligated into the pImp-1P backbone with T4 DNA ligase (New England Biolabs, NEB, Ispwitch, MA) overnight at 16°C.

Alanine-substituted LtrA mutant ORFs were transferred from the pImp-1P-derived vectors to the donor plasmid pACD2X for *in vivo* use. The segment of the LtrA ORFs containing the alanine or histidine substituted codon (GCA or CAT) from the pImp-1P-derived vectors was amplified by PCR. The primers used, LtrA551F (5'-GAAAATGAGCCAATTGATTTATAAATTTCTAAAAG-3') and LtrA1797R (5'-GGTGCTCGAGATATCTCACTTGTGTTTATGAATCACGTGACG-3), allow insertion into pACD2X between the MfeI and XhoI restriction sites.

### 2.5.2 Reconstitution of RNPs

Ll.LtrB intron RNPs were reconstituted using purified recombinant LtrA protein and *in vitro* synthesized Ll.LtrB-ΔORF lariat RNA using a modification of a previously described method (Saldanha et al. 1999). Wild-type LtrA protein was expressed from pImp-1P in *E. coli* BL21(DE3) grown in LB media containing 50 μg/ml ampicillin. pImp-1P expresses protein with a C-terminal intein-linked chitin-binding domain, which allows the LtrA to be purified via chitin-affinity column chromatography, using a slight modification of a previous protocol (Saldanha et al. 1999). Here, prior to loading the cell lysate to the chitin-affinity column, nucleic acid was precipitated by adding polyethyleneimine (PEI) dropwise to 0.1% while stirring at 4°C for 1 h, followed by centrifugation at 4000 X G for 30 min at 4°C (Avanti JA-I Centrifuge, JA-14 rotor, Beckman-Coulter, Fullerton, CA). Protein concentrations were determined by Bradford assay with LtrA protein whose concentration was previously determined by $A_{280}$ as a standard. The LtrA protein preparations were >98% pure based on SDS-polyacrylamide gels stained with Coomassie blue. Mutant LtrA proteins were expressed and purified as described for wild-type LtrA.

The 793-nucleotide Ll.LtrB-ΔORF precursor RNA, consisting of the 750-nucleotide Ll.LtrB-ΔORF intron flanked by the 25-nucleotide 5'-exon sequence and 18-nucleotide 3'-exon sequence, was transcribed from a BamHI-linearized pDR2C plasmid template with phage T7 polymerase, using a MEGAscript kit (Ambion, Austin, TX). The transcription reaction was terminated by phenol-chloroform-isoamyl alcohol (25:24:1. v:v:v; phenol-CIA, Ambion) extraction. Salt and free nucleotides were removed by

ethanol precipitation followed by a 70% ethanol wash. Ll.LtrB-ΔORF intron lariat was generated by self-splicing 3 nmol precursor RNA in 1 ml of 1.25 M $NH_4Cl$, 50 mM $MgCl_2$, and 40 mM Tris-HCl (pH 7.5) at 42°C. Prior to self-splicing, the precursor RNA was renatured in 1.25 M $NH_4Cl$, 5 mM $MgCl_2$, and 40 mM Tris-HCl (pH 7.5) by heating to 55°C and slowly cooling to 42°C. Self-splicing was initiated by adding $MgCl_2$ to 50 mM. The reaction mixture was incubated at 42°C for 3 h and then cooled to 30°C (Saldanha et al. 1999; San Filippo and Lambowitz 2002).

RNPs were reconstituted by adding the self-spliced RNA and 9 mmol purified LtrA to reconstitution medium pre-heated to 30°C and adjusting to 20 ml final volume with concentrations of 500 mM NaCl, 25 mM $NH_4Cl$, 5 mM $MgCl_2$, and 50 mM Tris-HCl (pH 7.5). The reconstitution reaction was incubated at 30°C for 90 min and the RNPs were isolated by ultracentrifugation in a Beckman Ti50.2 rotor (Beckman-Coulter; 40,000 rpm for ≥12 h at 4°C). The RNP pellet was resuspended in 300 µl of 10 mM $MgCl_2$, 10 mM KCl, and 50 mM Tris-HCl (pH 7.5), and stored at -80°C. Final concentrations of resuspended RNPs ranged from 1 µg/µl to 2 µg/µl. RNPs were reconstituted with mutant LtrA proteins and in-vitro synthesized Ll.LtrB-ΔORF lariat RNA as described above for wild-type RNPs.

### 2.5.3 DNA substrates for cross-linking

DNA substrates, corresponding to the wild-type Ll.LtrB intron integration target site from positions -30 to +15 (see Figure 1.2 for sequence), used for preparative-scale cross-linking, were synthesized (Keck Biotechnology Resource Laboratory, Yale University, New Haven, CT) with site-specific substitutions of photoactive nucleotide analogues. The

4-S-dT+5 and 4-S-dT-23 DNA substrates contain a 4-thio-deoxythymine (4-S-dT) (Glen Research Corp., Sterling, VA) in place of the wild-type thymine at top strand positions +5 or -23, respectively.

For the DNA substrate used for analytical-scale cross-linking, a single-stranded oligonucleotide, corresponding to top strand positions +5 to +15, with 4-S-dT at the +5 position was synthesized (Keck Biotechnology Resource Laboratory, Yale University, New Haven, CT) and 5'-end labeled with [$\gamma$-$^{32}$P]ATP (PerkinElmer, Waltham, MA) using T4 polynucleotide kinase (NEB). The labeled oligonucleotide and an equimolar concentration of DNA target top strand oligonucleotide from positions -46 to +5 were annealed to an equimolar amount of DNA target bottom-strand oligonucleotide from positions -46 to +15 (Integrated DNA Technologies, Coralville, IA). The two top-strand oligonucleotides were then ligated with T4 DNA ligase (NEB) and phenol-CIA extracted. The substrate was purified in a 2% agarose gel to remove single stranded oligonucleotides and free nucleotides (PureLink Gel Extraction Kit, Invitrogen, Carlsbad, CA).

### 2.5.4 Cross-linking and analysis

For preparative-scale cross-linking of the LtrA protein to the DNA substrate, 0.5 nmol of reconstituted RNPs and 5 nmol of DNA substrate were incubated for 15 min at 37°C in 4.5 ml reverse-splicing buffer containing 10 mM $MgCl_2$, 10 mM KCl, and 50 mM Tris-HCl (pH 7.5). The reaction mixture was spotted in 50 µl aliquots onto a silane-treated glass plate over ice for UV irradiation. The spots were irradiated over ice for 45 min with a 365-nm UV lamp (Spectroline, Westbury, NY).

Following irradiation, the spots were collected and combined. $CaCl_2$ was added to 5 mM and the reaction mixture was incubated with micrococcal nuclease (NEB) and Turbo DNase (Ambion) at 37°C for 1 h. The protein was precipitated by adding four volumes of -20°C, 10% (w/v) trichloroacetic acid (TCA) in acetone, incubating overnight at -80°C, and centrifuging at 10,000 rpm for 1 h at 4°C. The pellet was washed in 300 µl -20°C acetone to remove residual TCA, and air-dried to remove residual acetone.

For trypsin digestion, the pellet was resolubilized by adding 20 µl 0.2% ProteasMAX surfactant (Promega, Madison, WI) in 50 mM $NH_4HCO_3$ and 15 µl 8 M urea and vortexing on medium speed at room temperature until completely resuspended. The mixture was diluted to 100 µl with 50 mM $NH_4HCO_3$. The LtrA protein was then reduced by adding dithiothreitol (DTT) to 5 mM and incubating at 56°C for 20 min. 1 µl of 1% ProteasMAX surfactant in 50 mM $NH_4HCO_3$ and 1.8 µg sequencing-grade Trypsin Gold (Promega) were added, and the reaction was incubated overnight at 37°C.

The mixture of peptides from the protease digest was enriched for the cross-linked peptide using Titansphere Phos-TiO spin columns according to the manufacturer's protocol (GL Sciences Inc. Tokyo, Japan) (Thingholm et al. 2006). The enriched peptide cross-link mixture was analyzed by electrospray-ionization (ESI) tandem mass spectrometry in negative-ion mode (Thermo LTQ mass spectrometer, Thermo Electron Corp., San Jose, CA). In samples introduced by direct infusion, a Harvard syringe pump (Holliston, MA) was used at a flow rate of 3 µl/min. For other samples, as indicated, in-line reverse-phase HPLC (RP-HPLC) using a Dionex Ultimate 3000 capillary HPLC system (Sunnyvale, CA) with a UV detector (VWD-3100) set at 260 nm and a 300 µm x

15 cm Dionex Acclaim PepMap 100 C18 column, was required to separate cross-linked peptides from background contaminants before introduction by ESI. The ESI source was operated in the negative-ion mode with an electrospray voltage of 3.5 kV and a heated capillary temperature of 90°C.

For analytical-scale cross-linking, the radiolabeled DNA substrate described above was used so that cross-linking could be qualitatively observed by the transfer of the radiolabel to the protein. 5 pmol of reconstituted RNPs were cross-linked with 50 pmol of DNA substrate containing a radiolabeled photoactive nucleotide analogue in 50 μl reverse-splicing buffer, Turbo DNase digested, and acetone precipitated as described above. The samples were resolubilized in 25 μl 1X Laemmli SDS-polyacrylamide gel loading buffer, heated to 95°C for 5 min, and centrifuged at 15,000 rpm for 1 min. 10 μl of each sample was loaded onto a 10% Tris-HCl Mini-gel (BioRad. Hercules CA) and run at 25 mA. The gel was soaked in isopropanol/glycerol/acetic acid (25:20:10, v/v/v) solution for 1 h before drying under vacuum at 80°C overnight. The dried gel was visualized by autoradiography.

### 2.5.5 Intron mobility assays

Intron mobility was assayed by using a genetic system in which a modified Ll.LtrB intron containing a phage T7 promoter inserted near its 3'-end inserts into a target site upstream of a promoterless $tet^R$ gene, thereby activating the expression of that gene (Guo et al. 2000; Karberg et al. 2001). The mobility assays used intron donor plasmid pACD2X encoding the Ll.LtrB-ΔORF intron with wild-type or mutant LtrA proteins, and the recipient plasmid pBRR3–ltrB, containing wild-type or mutant Ll.LtrB intron target

sites, cloned upstream of the promoterless *tet*$^R$ gene. The donor plasmid was electroporated or chemically transformed into *E. coli* HMS174(DE3) harboring the recipient plasmid, and cells were grown overnight at 37°C in LB medium containing chloramphenicol and ampicillin. The cells were then incubated with 100 $\mu$M IPTG for 1 h at 37°C, diluted, and plated onto LB medium containing ampicillin without or with tetracycline to select Tet$^R$ + Amp$^R$ colonies in which the intron had inserted into the target site (Guo et al. 2000). Mobility efficiencies were determined as the ratio of Tet$^R$ + Amp$^R$ / Amp$^R$ colonies. Mobility assays with site-specific amino acid variants of LtrA at DNA recognition sites as determined by cross-linking were carried out as described above.

Mobility assays used to test for replication orientation bias were conducted as described previously (Zhong and Lambowitz 2003). The recipient plasmids, pBRR3A-ltrB and pBRR3B-ltrB, contain the Ll.LtrB intron target site/*tet*$^R$ cassette cloned in opposite orientations relative to the direction of DNA replication. The donor plasmid was electroporated or chemically transformed into *E. coli* HMS174(DE3) harboring the recipient plasmid, and cells were grown overnight at 37°C in LB medium containing chloramphenicol and ampicillin. The cells were then incubated with 100 $\mu$M IPTG for 1 h at 37°C, diluted, and plated onto LB medium containing ampicillin with or without tetracycline to select Tet$^R$ + Amp$^R$ colonies in which the intron had inserted into the target site (Guo et al. 2000). Mobility efficiencies were again determined as the ratio of Tet$^R$ + Amp$^R$ / Amp$^R$ colonies.

### 2.5.6 Biochemical assays

DNA reverse splicing and endonuclease assays were performed with an internally-labeled 129-bp DNA substrate generated by PCR in the presence of [α-$^{32}$P]dTTP (3,000 Ci/mmol; Perkin Elmer, Waltham, MA) (San Filippo and Lambowitz 2002). The DNA (1.25 nM, 56,000 cpm) was incubated with RNPs (12.5 nM) at 37°C for 5 min in 20 μl of reverse splicing medium containing 10 mM KCl, 10 mM $MgCl_2$, and 50 mM Tris-HCl (pH 7.5). The reactions were terminated by phenol-CIA extraction and ethanol precipitated. The products were analyzed by electrophoresis in a denaturing 6% polyacrylamide gel, which was dried and imaged by autoradiography. The incubation time of 5 min is within the linear range, which was determined by performing a time course DNA reverse splicing and endonuclease assay with product formation measured at 0, 1, 5, 10 15, 20, 30, 45, 60, 120, and 360 min.

RT assays were performed by incubating 10 μM protein pre-heated to 30°C with pre-heated 10 μM poly-rA oligonucleotide (~350 nucleotide, Sigma-Aldrich, St. Louis, MO) pre-annealed to 10 μM 42-nucleotide deoxythymine oligonucleotide (Integrated DNA Technologies) in 50 μl 20 mM Tris-HCl at pH 7.5, 75 mM KCl, 10 mM $MgCl_2$, and 33 nM [α-$^{32}$P]dTTP (3,000 Ci/mmol; Perkin Elmer) at 30°C for 5 min. The reaction was stopped by adding 50 μl 0.5 M EDTA on ice. 5 μl of the reaction mixture was spotted onto DE81 Whatman chromatography paper (Sigma-Aldrich). The paper was dried and washed for 5 min in 1X saline-sodium citrate (SSC) buffer three times. The paper was then air-died and imaged by autoradiography. The incubation time of 5 min is within the linear range which was determined by performing a time course protein RT

assay with [α-$^{32}$P]dTTP-incorporation measured at 0, 10 s, 20 s, 30 s, 1 min, 2 min, 5 min, 10 min, 30 min, and 1 h.

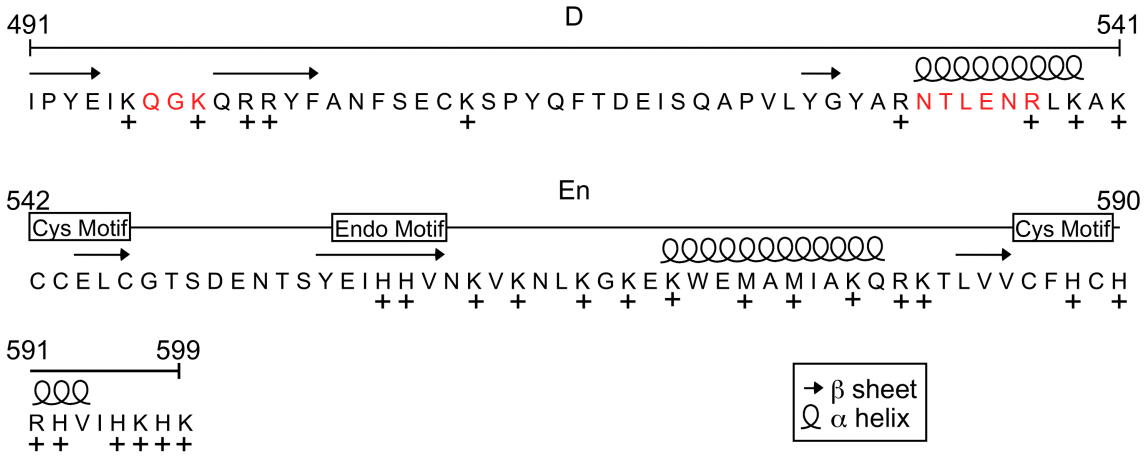


Figure 2.1: The LtrA C-terminal domains

The LtrA DNA-binding (D) domain and endonuclease (En) domain primary sequence is shown with α-helices and β-sheets predicted by the JPred3 program (http://www.compbio.dundee.ac.uk/www-jpred/) indicated above the sequence. A "+" indicates a positively charged amino acid residue. The conserved cysteine residue motifs (Cys Motif) and the endonuclease motif (Endo Motif) are labeled above the sequence (San Filippo and Lambowitz 2002; Blocker et al. 2005). The two peptides identified by UV-cross-linking as DNA target site contacts are in red. The peptide QGK (amino acids 497-499) contacts the T+5 position of the DNA target site and is located at the N-terminus of DNA-binding domain in a cluster of positively charged amino acid residues. The peptide NTLENR (amino acids 532-537) contacts the T-23 position of the DNA target site and is located within a predicted α-helix near the C-terminus of DNA-binding domain.

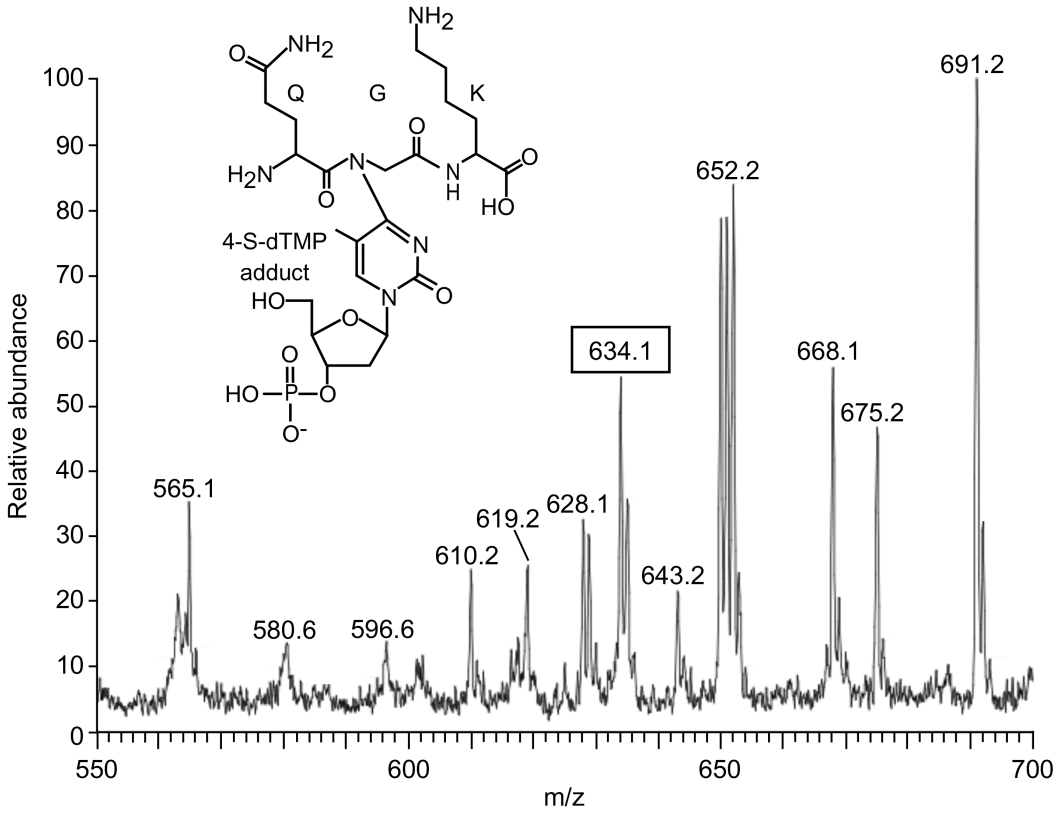


Figure 2.2: Full mass spectrum of the 4-S-dT+5-LtrA cross-link digest

The full negative-ion mode mass spectrum (ESI-ITMS) of the enriched 4-S-dT+5 cross-link mixture is shown with the proposed cross-link structure inset. Detection of the singly-charged cross-linked peptide at 634 m/z (boxed) did not require separation from other background species by LC. The cross-link bond is shown at the peptide backbone nitrogen of G498 for illustrative purposes. The exact position of cannot be determined based on the MS data (see Figure 2.3).

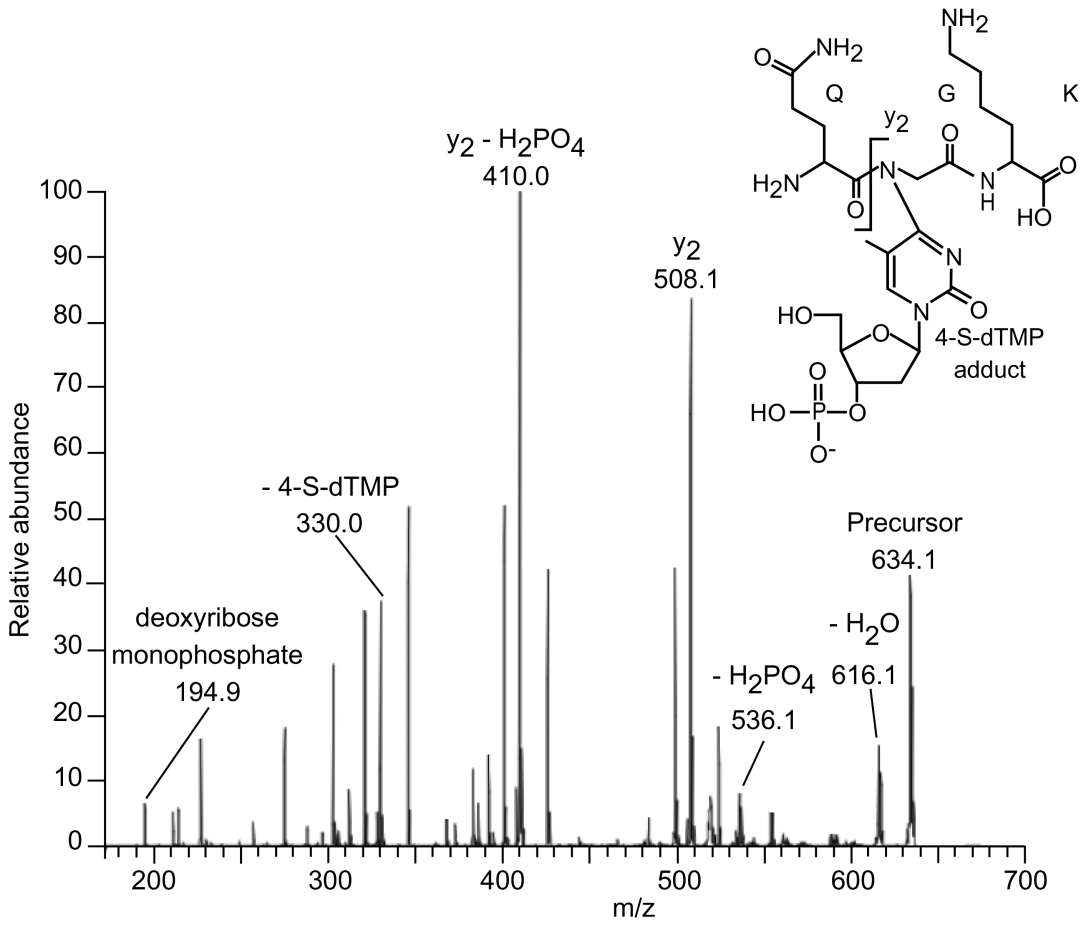


Figure 2.3: Tandem mass spectrum of the 4-S-dT+5-LtrA cross-link peptide

The CID tandem mass spectrum (ESI-CID-ITMS/MS) of the 634 m/z precursor ion contains fragments expected to be detected in the negative-ion mode. Ions resulting from water neutral loss from the precursor (616 m/z), as well as a phosphate loss from the precursor (536 m/z) are indicated. The $y_2$ fragment ion plus the mass of the 4-S-dTMP adduct (508 m/z) localizes the cross-link to G498 or K499. The cross-link bond is shown at the backbone nitrogen of G498, but the actual cross-link could be at any atom in G498 or K499. Because the $y_2$ ion includes the cross-link, it is subject to phosphate loss (410 m/z). The 4-S-dTMP fragment is also observed (330 m/z), indicating fragmentation at the cross-link bond. Additionally, detection of the deoxyribose monophosphate (195 m/z) indicates cleavage of the glycosidic bond between the deoxyribose monophosphate and the 4-S-dT base.

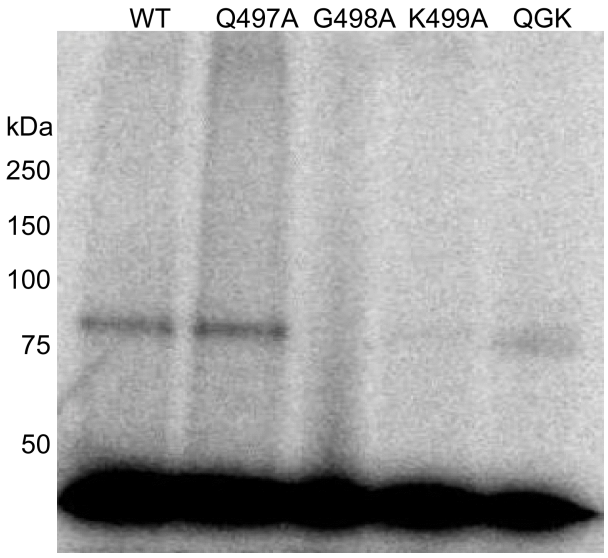


Figure 2.4: Analytical cross-linking with alanine-substituted LtrA (positions 497-499)

RNPs containing wild-type and mutant LtrA proteins were incubated with $^{32}$P-labeled 4-S-dT+5 DNA substrate and cross-linked with 365-nm light. Nucleic acids were then digested with DNase I and the protein was precipitated with acetone/TCA and analyzed by electrophoresis in a 0.1% SDS-7.5% polyacrylamide gel. The positions of prestained protein size makers (Kaleidoscope Prestained Standard, Bio-Rad)) are indicated to the left of the gel but the markers are not shown. LtrA is approximately 80 kDa. The Q497A LtrA mutant cross-links at wild-type levels. Cross-linking is significantly reduced with the QGK triple substitution LtrA and is virtually abolished with G498A LtrA and K499A LtrA.

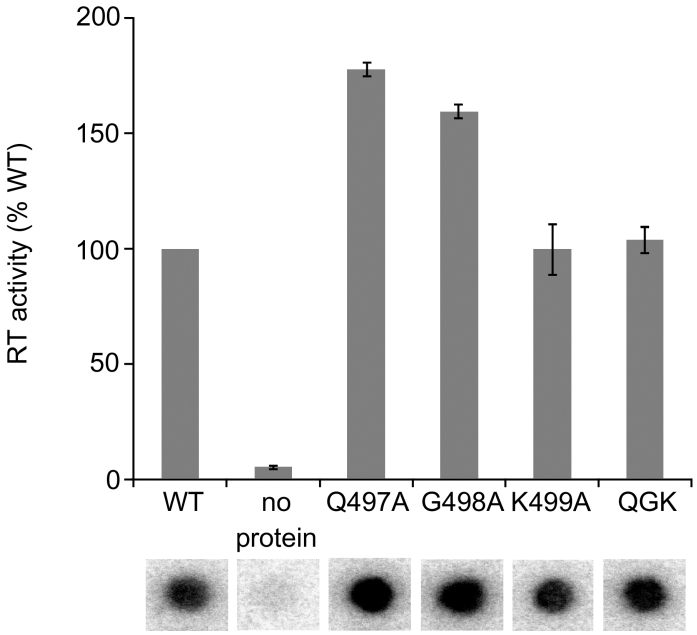


Figure 2.5: RT assays of alanine-substituted LtrA (positions 497-499)

Purified wild-type and mutant LtrA proteins (10 μM) were incubated with the artificial template-primer substrate poly(rA)/oligo(dT)$_{45}$ and polymerization of [α-$^{32}$P]dTTP was quantified by binding to DEAE paper. Incubations were at 30°C for 5 min, which was confirmed to be in the linear range by time course experiments. The error bars indicate standard deviation for at least three experiments. The RT activity of all four mutants is at or above the level of wild-type LtrA.

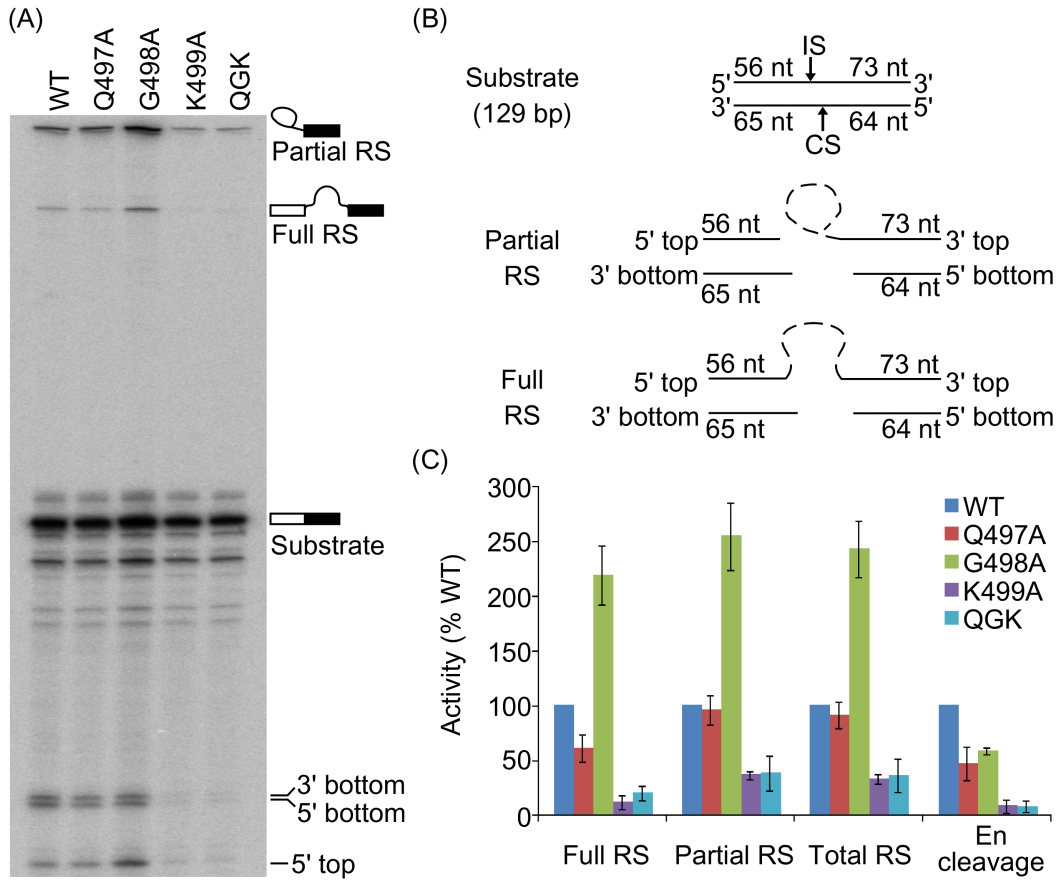


Figure 2.6: Reverse splicing and DNA endonuclease assays with alanine-substituted LtrA (positions 497-499)

(A) Ll.LtrB intron RNPs (5 $\mu$M, based on RNA $A_{260}$) containing either wild-type or mutant LtrA proteins were incubated with a 129-bp $^{32}$P-labeled DNA oligonucleotide substrate containing the Ll.LtrB intron insertion site (0.5 $\mu$M) at 37°C for 5 min, which was confirmed to be in the linear range by time course experiments, and the products were ethanol precipitated, phenol-CIA extracted, and analyzed in a denaturing 10% polyacrylamide gel. Products are indicated to the right of the gel. "Partial RS" refers to the partially reverse spliced intermediate and "Full RS" refers to the fully reverse spliced product. (B) In the schematic, the DNA substrate is represented as a solid line with the intron insertion site (IS) and bottom-strand cleavage site (CS) indicated. DNA fragment sizes, in number of nucleotides (nt), are denoted above each fragment. The partially reverse spliced intermediate and fully reverse spliced product are shown with intron RNA represented by a dashed line. (C) The bar graphs show biochemical activities relative to wild-type LtrA. “Total RS” refers to the combined amount of partially and fully reverse spliced products. The error bars indicate standard deviation for at least three experiments.

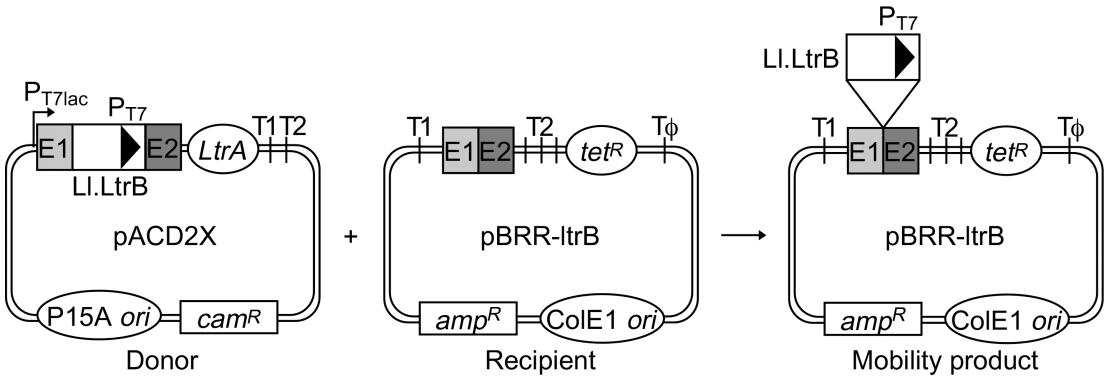


Figure 2.7: The *E. coli* two-plasmid mobility assay

The intron donor plasmid pACD2X carries a $cam^R$ marker and uses a T7lac promoter ($P_{T7lac}$) to express a 940-nt Ll.LtrB intron, flanked by the 5'-exon (E1) and 3'-exon (E2), with a phage T7 promoter ($P_{T7}$) inserted in DIV of the intron. The LtrA protein is expressed from a position just downstream of the 3'-exon. The recipient plasmid pBRR-ltrB carries an $amp^R$ marker and contains the Ll.LtrB intron target site (ligated 5'- and 3'-exon) cloned upstream of a promoterless $tet^R$ gene. The integration of the intron into the target site inserts $P_{T7}$ upstream of the $tet^R$ gene to activate expression, yielding $Amp^R$ + $Tet^R$ colonies. T1, T2, and Tϕ are *E. coli rrnB* T1, T2 and phage T7 transcription terminators, respectively.

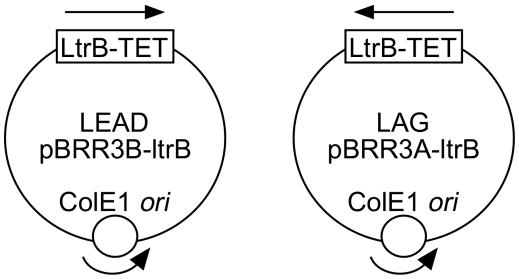


Figure 2.8: Schematic of recipient plasmids used for replication orientation bias mobility assays

pBRR3B-ltrB and pBRR3A-ltrB, denoted LEAD and LAG, respectively, depending on whether the nascent lagging or leading DNA strands could be used as the primer for reverse transcription of the inserted intron RNA, contain the target site/*tet*$^{R}$ cassette cloned in opposite orientations.

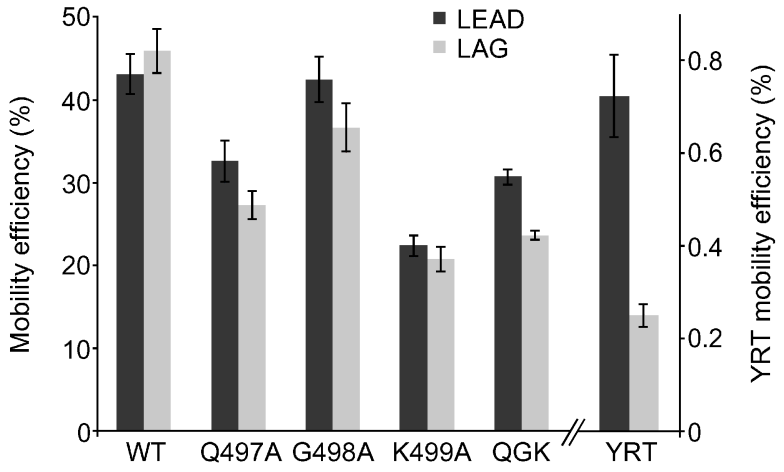


Figure 2.9: Mobility assays with alanine-substituted LtrA (positions 497-499)

Mobility assays were performed as described in Figure 2.7 using the LEAD and LAG recipient plasmids in Figure 2.8, with 100 μM IPTG induction at 37°C for 1 h. All bar graphs show mobility efficiencies, defined as the ratio ($Amp^R$ + $Tet^R$ colonies) / ($Amp^R$ colonies), expressed as a percentage. The left side axis is for wild-type, Q497A, G498A, K499A, and QGK LtrA. The right side axis is for YRT LtrA. Error bars indicate standard error for at least 3 experiments.

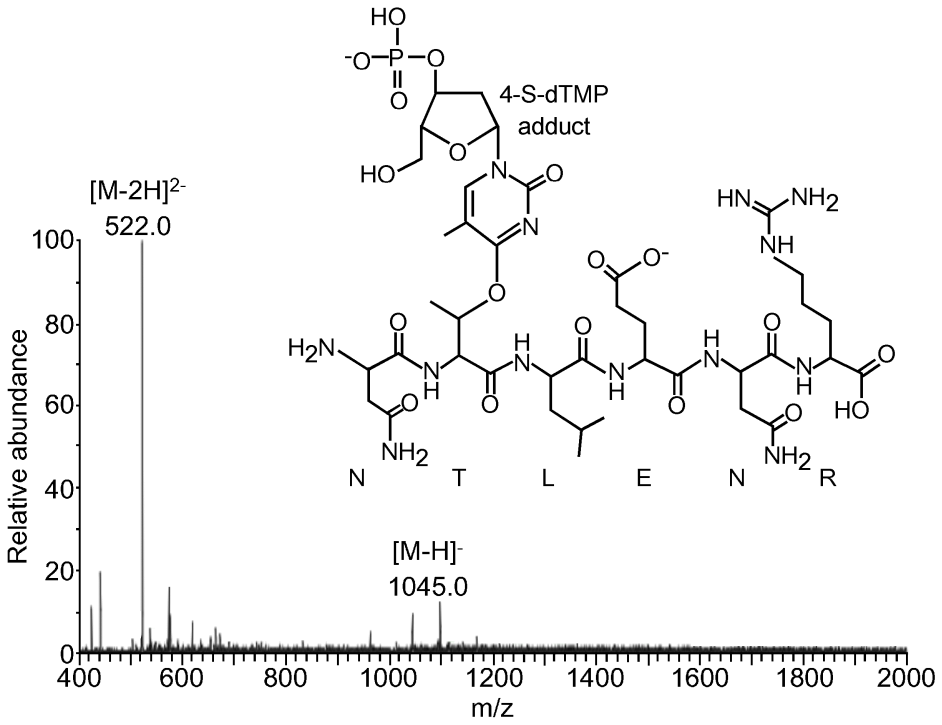


Figure 2.10: Full mass spectrum of the 4-S-dT-23-LtrA cross-link digest

The full negative-ion mode mass spectrum (LC-ESI-ITMS; LC retention time = 20.5 min) of the 4-S-dT-23 cross-link is shown with the proposed peptide cross-link structure. The species is present in both singly- (m/z = 1,045) and doubly-charged (m/z = 522) states.

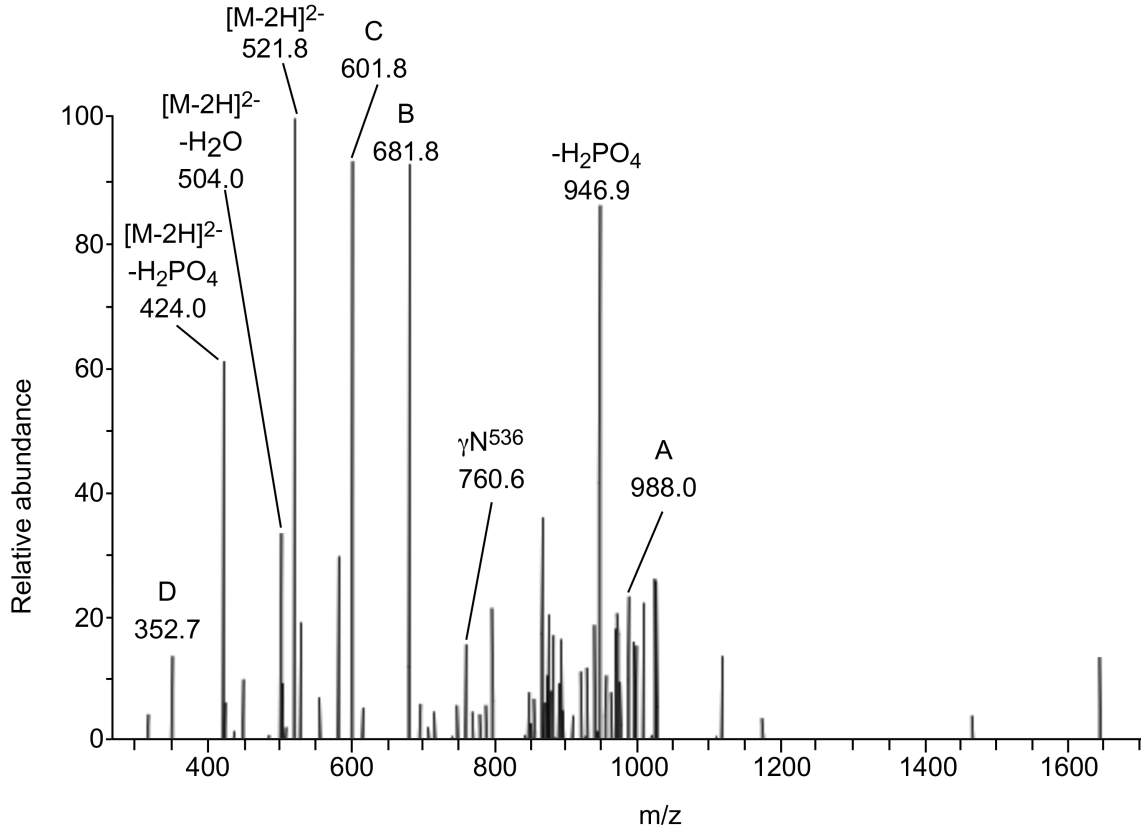


Figure 2.11: Tandem mass spectrum of the 4-S-dT+5-LtrA cross-link peptide

The CID tandem mass spectrum (LC-ESI-CID-ITMS/MS) of the 1,045 m/z precursor ion shows the doubly-charged $[M-H]^{2-}$ precursor ion (522 m/z), along with the corresponding water neutral loss ion (504 m/z) and the phosphate loss ion (424 m/z). The phosphate loss ion (947 m/z) of the singly-charged species is also observed. γ cleavage of the peptide backbone, mediated by the N536 side chain (761 m/z), is observed, but no backbone cleavage independent of side-chain chemistry occurs, a phenomenon that has been seen in previous studies of similarly sized peptides in negative-ion mode (Bowie et al. 2002). Peaks labeled A-D are produced by a series of side-chain eliminations, including the 4-S-dTMP adduct, and typical neutral loss from the C- and N-termini of the precursor peptide (side chain or other source of loss indicated in parenthesis): peak A $NH_3$ (N532 or N536), $H_2O$ (E535), $H_2O$ (C-terminus); peak B $NH_3$ (N532 or N536), acetone-4-S-dTMP (T533); peak C $NH_3$ (N532 or N536), acetone-4-S-dTMP (T533), $H_2O$ (E535), $H_2O$ (C-terminus), acetonitrile (R537). Peaks B and C both result from the elimination of the T533 side chain, however, rather than the typical neutral loss of acetone ($CH_3$-CHO) from threonine (Bowie et al. 2002), the lost mass includes the entire 4-S-dTMP adduct, indicating the location of the cross-link bond. Peak D is the other product of the T533 side chain elimination, the acetone-4-S-dTMP adduct.

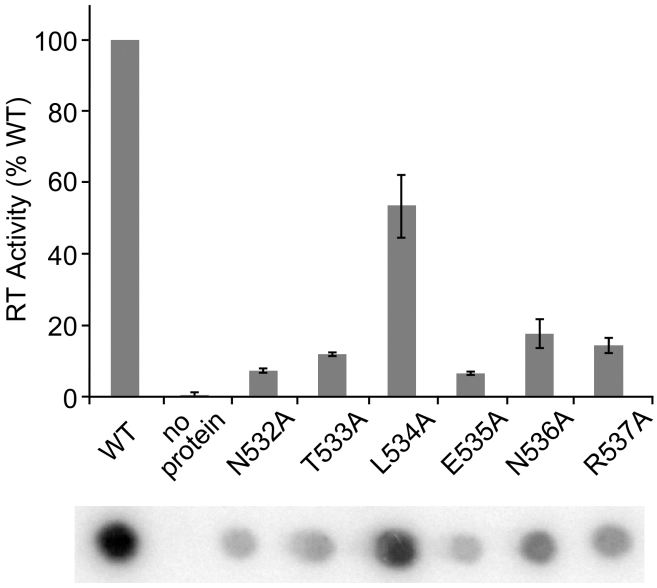


Figure 2.12: RT assays of alanine-substituted LtrA (positions 532-537)

Purified wild-type and mutant LtrA proteins were incubated with the artificial template-primer substrate poly(rA)/oligo(dT)$_{45}$ and polymerization of [$\alpha$-$^{32}$P]dTTP was quantified by binding to DEAE paper. Incubations were at 30°C for 5 min, which was confirmed to be in the linear range by time course experiments. The error bars indicate standard deviation for at least three experiments.

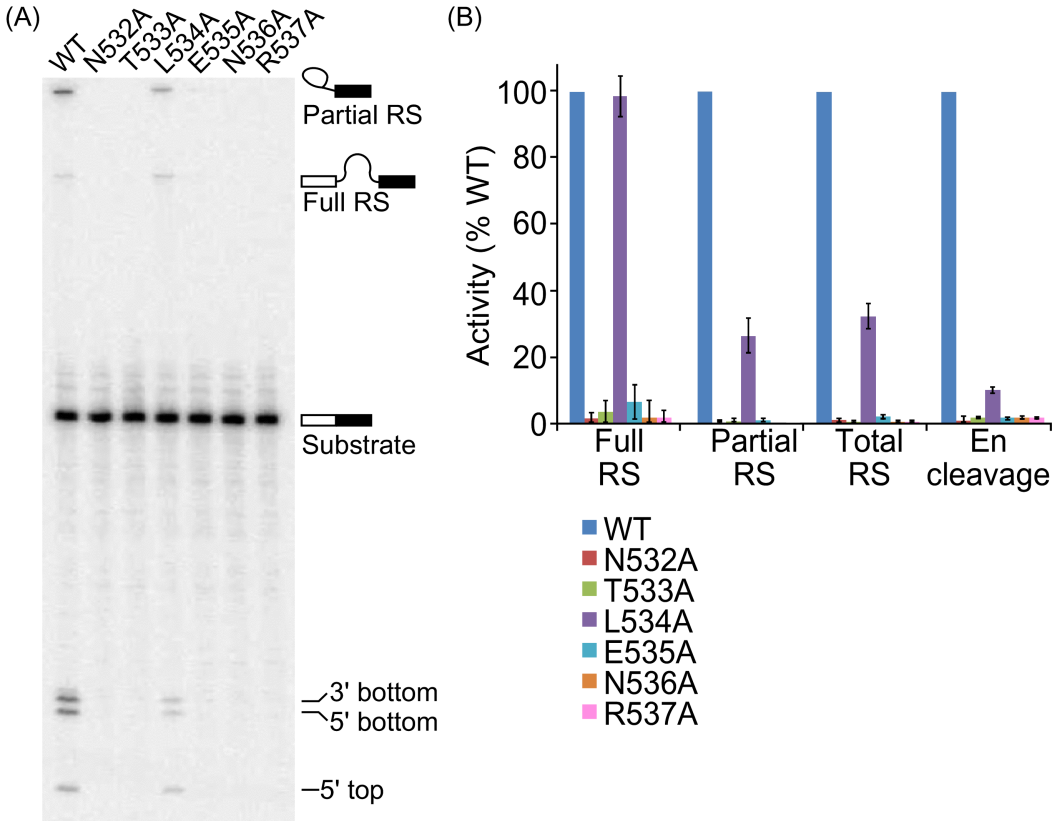


Figure 2.13: Reverse splicing and DNA endonuclease assays with alanine-substituted LtrA (positions 532-537)

(A) Ll.LtrB intron RNPs (5 $\mu$M) containing either wild-type or mutant LtrA proteins were incubated with a 129-bp $^{32}$P-labeled DNA oligonucleotide substrate (0.5 $\mu$M) containing the Ll.LtrB intron insertion site at 37°C for 5 min, which was confirmed to be in the linear range by time course experiments, and the products were analyzed on a denaturing 10% polyacrylamide gel. Products, as described in Figure 2.6B, are indicated to the right of the gel. (B) The bar graphs show biochemical activities relative to wild-type LtrA. "Total RS" refers to the combined amount of partially and fully reverse spliced products. The error bars indicate standard deviation for at least three experiments.

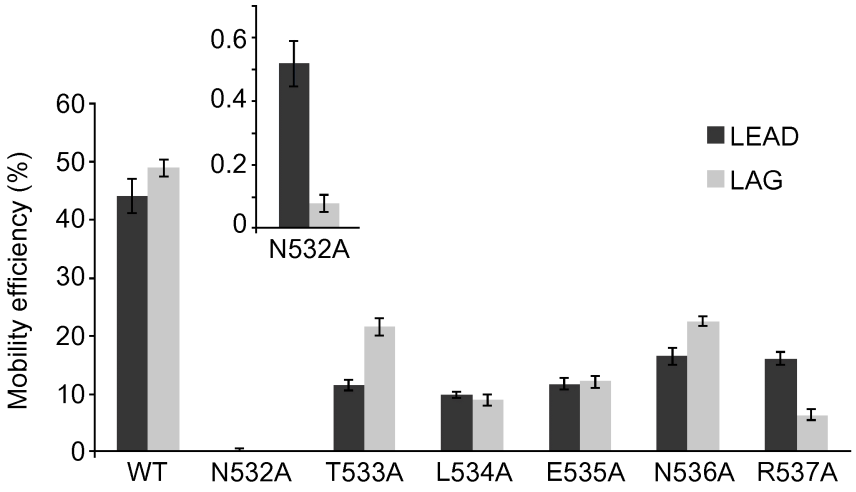


Figure 2.14: Mobility assays with alanine-substituted LtrA (positions 532-537)

Intron mobility assays with wild-type LtrA, or the N532A, T533A, L534A, E535A, N536A, or R537A mutants were performed as described in Figure 2.7, using the LEAD and LAG recipient plasmids described in Figure 2.8. All bar graphs show mobility efficiencies, defined as the ratio ($Amp^R$ + $Tet^R$ colonies) / ($Amp^R$ colonies), expressed as a percentage. The inset graphs show the mobility efficiencies of N532A with an expanded scale. The error bars indicate standard error for at least three experiments.

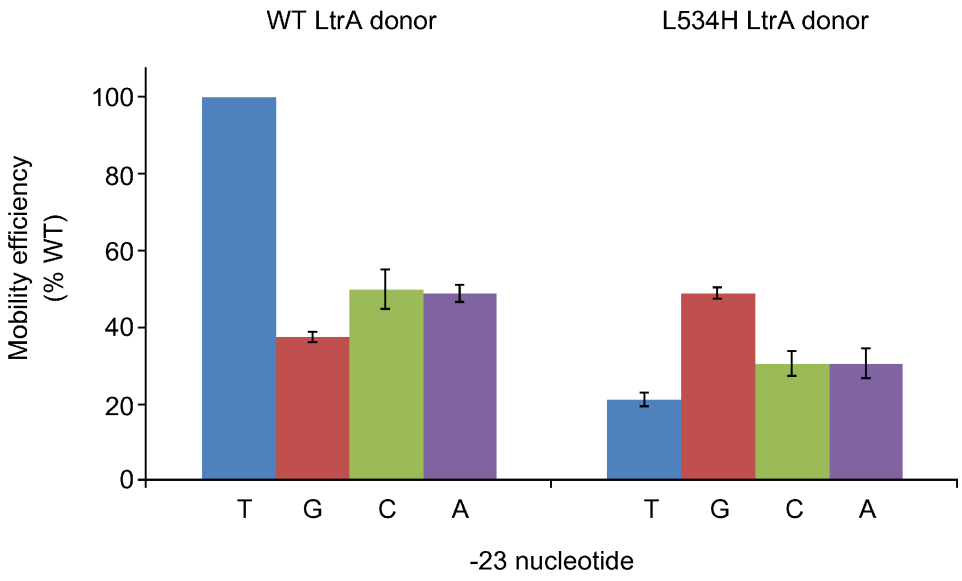


Figure 2.15: L534H LtrA specificity switch mobility assay

Mobility assays were performed as described in Figure 2.7 using a series of -23 position-mutant recipient plasmids to test for altered specificity. The recipient plasmids contained either T (WT), G, C, or A residues at the -23 position of the DNA target site. The bar graphs show mobility efficiency relative to wild-type LtrA with wild-type (T-23) pBRR-ltrB recipient. The error bars indicate standard error for at least three experiments.

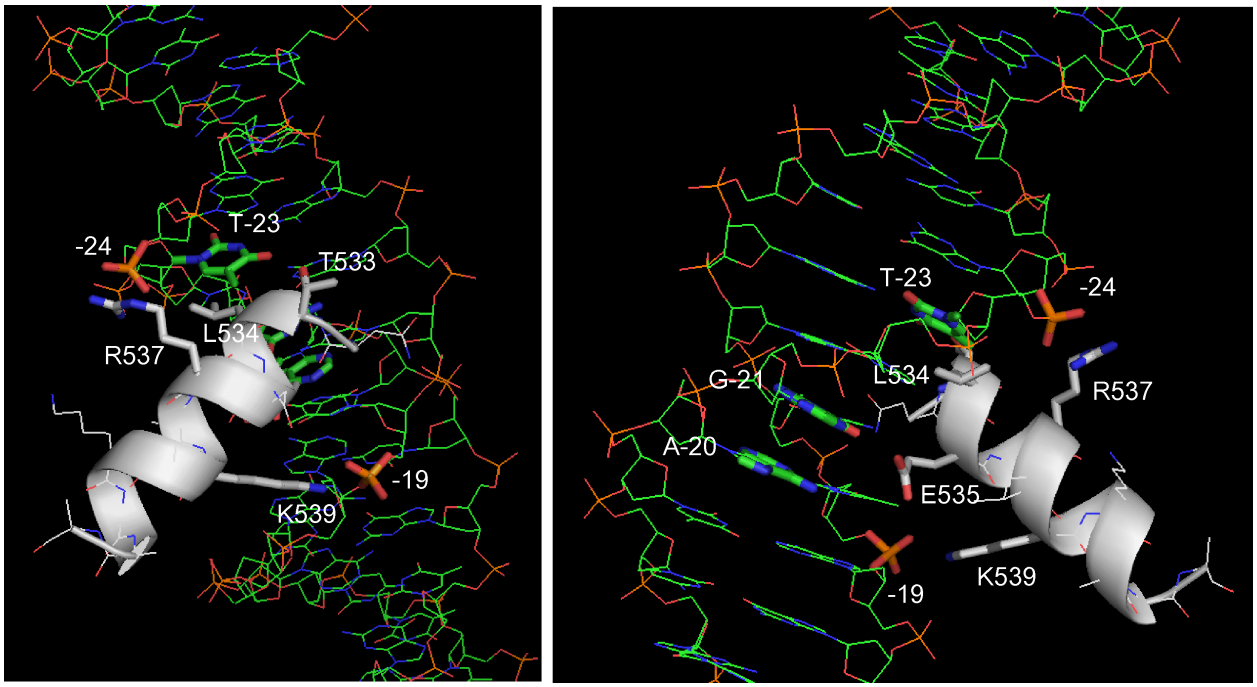


Figure 2.16: Model of LtrA DNA-binding domain α-helix interaction with the Ll.LtrB DNA target site

The conserved putative α-helix (white) at the C-terminal end of the LtrA DNA-binding domain is shown from two different perspectives docked to the major groove of the 5'-exon of the DNA target site. Key residues/chemical groups are highlighted. The α-helix is positioned to permit hydrogen bonding between the T533 side-chain hydroxyl and the O4 carbonyl of the thymine base at the -23 DNA position, in agreement with the DNA cross-linking data. The L534 side-chain is positioned to make hydrophobic contacts with the C5 methyl group of the thymine base, as implied by the L534H genetic switch mobility assay data and L534A mutational analysis. R537 and K539 are positioned to contact the -24 (top strand) and -19 (bottom strand) phosphate groups, respectively. E535 could potentially contact either the G-21 or A-20 bases through hydrogen bonding or water-mediated bonding depending on the side-chain rotamer in the actual structure.

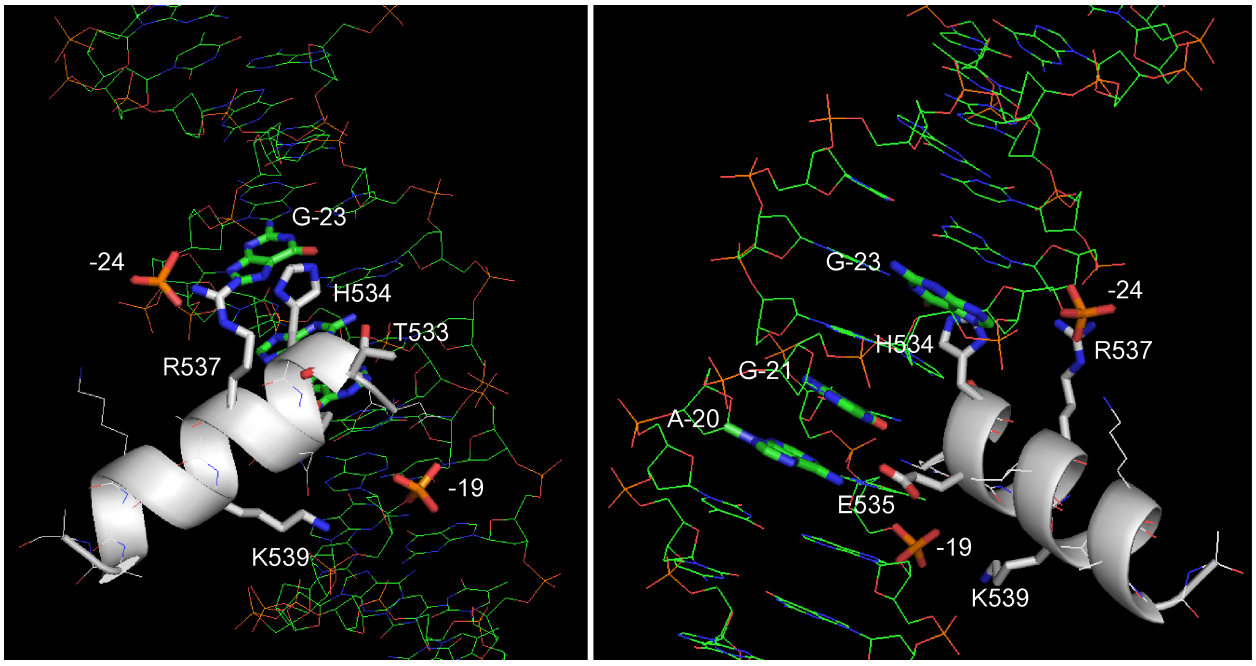


Figure 2.17: Model of L534H LtrA DNA-binding domain α-helix interaction with the Ll.LtrB T-23G mutant DNA target site

The conserved putative α-helix (white) at the C-terminal end of the LtrA DNA-binding domain is shown from two different perspectives docked to the major groove of the 5'-exon of the DNA target site with the T-23G mutation. Key residues/chemical groups are highlighted. The bulky histidine side-chain of L534H prevents T533 from contacting the -23 position of the DNA target site. Instead, the α-helix is slightly reoriented to accommodate the histidine, which hydrogen bonds to the guanine at the -23 position. Most other contacts described in Figure 2.16 may still be made but with imperfect geometries.

# Chapter 3: DNA bending during Ll.LtrB intron integration into DNA

The complex process of group II intron retrohoming requires the IEP to contact specific bases in the distal 5'-exon and the 3'-exon, possibly at the same time (Mohr et al. 2000; Singh and Lambowitz 2001). A three-dimensional homology model of the LtrA protein, formulated by threading the LtrA dimer onto X-ray crystal structures of the HIV-1 RT dimer, suggests that the DNA target sequence is too long for LtrA to simultaneously contact both exons unless the DNA is strongly bent in the complex (Blocker et al. 2005). These findings raise the possibility that the propensity of a DNA sequence to bend (the sequence's "bendability"), in addition to the sequence requirements detailed in Section 1.3 (Figure 1.2), might significantly influence the efficiency of integration at that DNA target site.

Here, I analyze the bending of the Ll.LtrB intron target sites by atomic force microscopy (AFM). The computational analysis of Ll.LtrB intron target site bendability by Dr. Jiri Perutka is summarized in Section 3.1. It concludes that Ll.LtrB DNA target sites have significantly higher predicted bendability than do random *E. coli* DNA sequences. This prefaces the collaborative work of Dr. James Noah and I, described in Sections 3.2 and 3.3, where AFM is used to show that the target DNA is bent at two progressively greater angles during the integration of group II intron RNPs, which reflect different stages of the retrohoming reaction pathway.

## 3.1 Computational evidence for DNA bendability as a DNA target site feature

An algorithm that used a probabilistic model based on the bendability of trinucleotides deduced from DNase I-digestion data (Brukner et al. 1995) calculated an average bendability profile for two sets of Ll.LtrB insertions sites in the *E. coli* genome, one consisting of 66 retrohoming sites used by an Ll.LtrB intron having randomized target-site recognition sequences (Zhong et al. 2003), and the other consisting of 33 retrotransposition sites used by the wild-type Ll.LtrB intron (Coros et al. 2005). Both test sets were culled from larger sets of sites by excluding those lacking T+5, which may have been utilized via En-independent pathways that do not require the same IEP interactions with the 3'-exon, as described in Section 2.2.3 (Zhong and Lambowitz 2003; Coros et al. 2005; Yao et al. 2005). The average bendability profiles for the tests sets were then compared with those for 100 different control sets, each consisting of the same number of randomly selected 45-bp *E. coli* genomic sequences (a total of 6600 and 3300 control sequences for the retrohoming and retrotransposition sites, respectively). A collection of target sites provides a better picture of bendability patterns than examination of individual target sites, including the wild-type site, in which the influence of bendability might be obscured by other features.

Figure 3.1 shows the bendability profiles for the retrohoming sites (red), retrotransposition sites (blue), the wild-type Ll.LtrB intron target site (dashed line), and 100 control sets for the retrohoming sites (light blue). The profiles were calculated for positions -30 to +10, using a 6-bp sliding window and a shift increment of one. The control sets for the retrotransposition sites are similar to those for the retrohoming sites and were omitted to simplify the plots. At a 5% significance level, both the retrohoming and retrotransposition sites show several regions with significantly higher bendability

than the control sites, albeit at somewhat different locations in the retrohoming and retrotransposition sites (6-bp segments beginning at positions -13 to -10 and -3 in the retrohoming sites and at positions -20 to -18 and -6 in the retrotransposition sites). A possible explanation for the different locations of the bendable regions is that the retrotransposition sites were obtained with the wild-type Ll.LtrB intron, while the retrohoming sites were obtained with an Ll.LtrB intron having randomized EBS2, EBS1, and δ sequences. Consequently, the retrotransposition sites are constrained to have target sequences that base pair with a single intron (wild-type Ll.LtrB), while the retrohoming sites are not similarly constrained, providing more latitude for the location of the bendable regions. This inference is supported by the finding that the bendability profile for the wild-type retrohoming site (dashed line), which must also base pair with the wild-type Ll.LtrB intron, parallels that for the retrotransposition sites. Notably, the bendability peaks for the wild-type site have significantly higher amplitudes than do those in the profiles calculated by averaging multiple retrohoming and retrotransposition sites (see difference in scale on the right). These higher amplitudes may reflect the importance of bendability in selection of the wild-type target site. The computational analysis indicates that both retrohoming and retrotransposition sites have regions of higher predicted bendability than do random *E. coli* DNA sequences. Additional computational comparisons of protein-induced deformability (Olson et al. 1998), propeller twist (el Hassan and Calladine 1996), and stacking energy (Ornstein 1978) also showed significant differences in the retrohoming and retrotransposition sites compared to the control sets,

reinforcing the conclusion that Ll.LtrB integration sites may have distinctive higher-order structural characteristics (Noah et al. 2006).

### 3.2 VISUALIZING RNP-DNA COMPLEXES BY AFM

Together with the HIV-1 RT-based LtrA model, the computational analysis of Ll.LtrB retrohoming and retrotransposition sites suggest that the bendability of DNA appears to be a significant factor in the suitability of DNA sequence for targeting by the Ll.LtrB intron RNP. In collaboration with Dr. James Noah, I used AFM to analyze bending of the Ll.LtrB intron DNA target site at different stages of the retrohoming reaction pathway using a series of mutant RNPs and altered DNA target sites. Dr. Noah designed the original set of DNA substrates used for the AFM experiments (Figure 3.2) and demonstrated their biochemical activity with Ll.LtrB intron RNPs (Figure 3.3, detailed description in Section 3.2.1). Additionally, he demonstrated that the RNP could be distinguished from the DNA substrate in Ll.LtrB RNP-DNA complexes visualized by AFM (Figure 3.4, detailed description in Section 3.2.2) and that the distribution of DNA bend angles is dependent on the incubation time prior to imaging (Figure 3.6, detailed description in Section 3.2.3).

The bulk of my contributions to this research are described in Section 3.3, but the imaging of free RNPs (Figure 3.5) and naked DNA substrates (Figure 3.7) discussed in Section 3.2 are also my work. In Section 3.3, I determined the role of IEP-DNA target site interactions in DNA bending at different stages of the retrohoming reaction using mutant RNPs and mutant or physically altered DNA target substrates. Additionally, DNA bending upon initiation of TPRT with the addition of dNTPs to the reaction was

examined. Together with previous findings, these results suggest a model in which RNPs bend the target DNA by maintaining initial IEP and base-pairing interactions with the distal 5'-exon, while engaging in subsequent interactions with the 3'-exon that successively position the scissile phosphate for bottom-strand cleavage at the En active site and then reposition the 3'-end of the cleaved bottom strand at the RT active site for initiation of reverse transcription.

### 3.2.1 Biochemical activity with DNA target substrates

A series of ~240-bp wild-type Ll.LtrB DNA target site substrates and mutant DNA substrates with unfavorable, non-wild-type nucleotide residues in different regions of the Ll.LtrB intron target site were used in the AFM experiments (Figure 3.2). They were tested for their ability to support intron reverse splicing and bottom-strand cleavage (Figure 3.3). In these experiments, which are very similar to those described in Sections 2.2.2 and 2.3.2, wild-type Ll.LtrB RNPs were incubated with internally $^{32}$P-labeled DNA substrates at 37°C for 5 min or 30 min.

As expected, the wild-type substrate (WT) supported both full and partial reverse splicing, and bottom-strand cleavage. Previously, it has been demonstrated that, with wild-type Ll.LtrB RNPs and the wild-type target site, reverse splicing products appear first, while bottom-strand cleavage products appear after a lag (Singh and Lambowitz 2001). In agreement with these findings, after 5 min, the reverse splicing products comprised 28% of the total radioactivity and bottom-strand cleavage products only 2%, while after 30 min, the reverse-splicing products comprised 52% and the bottom-strand cleavage products 9%.

The reactions with the remaining DNA substrates were done for 30 min. The L DNA substrate, which contains the wild-type target site near the 5'-end of the DNA molecule, reacted as efficiently as the WT DNA substrate, with the shift of the site resulting in different sizes for the bottom-strand cleavage products (Figure 3.3). The -30/+15 substrate, which has unfavorable non-wild-type target site nucleotide residues at all positions between -30 and +15, showed only a very low level of reverse splicing (< 1% that of the WT substrate), which is likely due to inefficiently utilized cryptic target sites (Mohr et al. 2000). Mutant DNA substrates -30/-13 and -12/+3, which have non-wild-type nucleotide residues in the distal 5'-exon and IBS/δ regions, respectively, also showed <1% of the reverse splicing activity of the WT substrate and no detectable bottom-strand cleavage. These results are in agreement with previous findings showing that mutations affecting either LtrA interactions with the distal 5'-exon region or EBS-IBS/δ-δ' base pairing inhibit both reverse splicing and bottom-strand cleavage (Mohr et al. 2000; Singh and Lambowitz 2001).

DNA substrate +4/+15, which has non-wild-type nucleotide residues at 3'-exon positions recognized by LtrA, supported efficient reverse splicing, but gave no detectable bottom-strand cleavage (Figure 3.3). In agreement with previous results (Matsuura et al. 1997; San Filippo and Lambowitz 2002), Ll.LtrB intron RNPs containing a C-terminally truncated IEP lacking the D and En domains (ΔD/En LtrA) showed no detectable reverse splicing or bottom-strand cleavage, while Ll.LtrB intron RNPs containing a truncated IEP lacking only the En domain (ΔEn LtrA) were still capable of reverse splicing, but were unable to carry out bottom-strand cleavage.

### 3.2.2 AFM with WT and L DNA target site substrates

Complexes formed between the RNPs and DNA substrates under the conditions of the reverse splicing and bottom-strand cleavage assays were imaged by AFM. Figure 3.4 shows representative complexes formed after incubating wild-type Ll.LtrB RNPs with the WT and L DNA substrates for 30 min at 37°C. The samples were air-dried and imaged in tapping mode. RNP aggregation and complex immobilization in solution prevented RNPs from being visualized in buffer. Only well-separated complexes with both DNA ends visible were used for analysis.

In the AFM images, the RNPs appeared as a bright dot located in the center or at the end of the WT and L DNA substrates, respectively. The shift in the position of the RNPs bound to the L substrate indicates that they specifically recognized, and bound the DNA target sequence and also that the RNP-DNA complexes were two-dimensionally fixed on the aminosilane-treated mica slide in their original position in the complex. Based on measurements of 100 molecules, the wild-type target DNA in complex with RNPs had a length of 85 ± 9 nm compared to 81 ± 11 nm for the free DNA (see below) and a calculated length of (3.3 Å x 234 bp =) 77.2 nm. The similar lengths of free and bound DNA indicate that the DNA is not substantially wrapped around the RNP in the complex.

At the resolution of the AFM, the RNPs bound to the DNA substrate appear roughly circular. Based on measurement of 100 molecules, RNPs bound to the wild-type DNA substrate had a diameter of 10 ± 1 nm and a height of 0.9 ± 0.2 nm, and the corresponding measurements for RNPs bound to the L DNA substrate were 11 ± 2 nm

and 1.1 ± 0.3 nm. The RNPs bound to the L DNA substrate in Figure 3.4 were among the larger RNPs observed (diameters of 11 and 12 nm and heights of 1.1 and 1.4 nm). AFM measurements of free RNPs under the same conditions gave a diameter of 11 ± 1 nm and height of 1.0 ± 0.4 nm, similar to the dimensions of the RNPs bound to the DNA (Figure 3.5). Although RNP damage caused by sample preparation or imaging cannot be excluded (Giro et al. 2004), only active RNPs containing both Ll.LtrB intron lariat RNA and the LtrA protein can react with the DNA substrate. RNA and protein by themselves do not bind DNA specifically (Singh and Lambowitz 2001). The relatively low height of the RNPs may reflect compaction by the tip, interaction of the RNPs with the mica substrate, dehydration, capillary forces during drying, or differences in compliance between the RNPs and mica substrate in tapping mode (Matsuura et al. 2006). It is possible that the RNPs are inherently non-globular and flattened in one dimension, with this flattening enhanced by dehydration, similar to the situation for *E. coli* small ribosomal subunits imaged in air (see Section 3.3)(Matsuura et al. 2006).

### 3.2.3 Time-dependent distribution of bend angles in the retrohoming pathway

The initial AFM imaging of complexes with the WT target site, performed with samples incubated for 30 min at 37°C, showed that the DNA is bent around the RNPs and that the bend angles have a distinctly bimodal distribution with peaks at ~75° and ~90°. To test whether the two DNA bend angles might reflect different stages of the reactions, 100 complexes with the WT target site after 5 and 30 min of incubation were compared (Figure 3.6). The lesser bend angle predominated in the 5 min samples (75 ± 6°, 83%; 88 ± 3°, 17%), while the greater bend predominated in the 30 min samples (75 ± 2°, 49%,

90° ± 5°, 51%). In the absence of RNPs, the same position in the WT target DNA showed a wide distribution of bend angles (Figure 3.7). The distribution of bend angles in the naked WT DNA substrate was skewed more toward larger bend angles than is the case for the -30/+15 mutant DNA (Figure 3.7), which possibly reflects the wild-type DNA target sequence's increased bendability even in the absence of RNPs (see Section 3.1).

## 3.3 DNA BENDING AT DIFFERENT STAGES OF THE RETROHOMING REACTION

### 3.3.1 AFM with mutant DNA target site substrates

As expected, complexes formed between wild-type Ll.LtrB RNPs and mutant DNA substrate -30/-13, which has unfavorable nucleotide residues in the distal 5'-exon region recognized by the IEP, or -12/+3, which has unfavorable nucleotide residues that disrupt EBS-IBS/δ-δ' base pairing to the intron RNA, were not detected. These findings agree with previous DNA-footprinting experiments, which showed that DNA substrates having unfavorable nucleotide residues in the distal 5'-exon region do not form stable complexes with the RNP (Singh and Lambowitz 2001).

The +4/+15 mutant DNA substrate, which has unfavorable nucleotides in the 3'-exon from position +4 to +15, could support reverse splicing, but not bottom-strand cleavage in the biochemical assays in Section 3.1, and, therefore, do form stable complexes with the RNP (Figure 3.3). Imaging of 100 complexes formed with the +4/+15 DNA substrate showed RNPs bound at the correct position, but with a wide range of DNA bend angles, similar to naked DNA (Figure 3.8). These findings indicate that initial DNA target site recognition and reverse splicing are not by themselves sufficient to

induce DNA bending, and that bending requires either bottom-strand cleavage itself or the additional LtrA interactions with the 3'-exon that lead to bottom-strand cleavage.

### 3.3.2 Imaging complexes with mutant RNPs

Similar conclusions were reached from AFM imaging of complexes formed with mutant RNPs containing C-terminally truncated LtrA proteins. RNPs reconstituted with the ΔD/En LtrA mutant, which lacks both the C-terminal DNA-binding and En domains, gave no detectable complexes, in agreement with DNA-footprinting experiments showing that these RNPs do not form stable complexes with the DNA substrate (Singh and Lambowitz 2001). RNPs reconstituted with the ΔEn LtrA mutant are capable of reverse splicing but not second-strand cleavage (Figure 3.3). As expected, these RNPs formed stable complexes detected by AFM, and analysis of 100 such complexes again showed RNPs bound at the correct position, but with a wide distribution of bend angles (Figure 3.8), reflecting either loss of required interactions with the 3'-exon that lead to bottom-strand cleavage or the lack of bottom-strand cleavage itself.

### 3.3.3 AFM of complexes with nicked DNA substrate

To distinguish whether DNA bending is caused by IEP interactions with the 3'-exon or by bottom-strand cleavage itself, I measured the DNA bend angles of wild-type RNP-bound DNA substrates with a pre-positioned nick at the bottom-strand cleavage site ("nicked" DNA substrates) (Figure 3.9). One nicked DNA substrate contained the wild-type DNA target sequence, and the other contained the 3'-exon +4/+15 mutations that block interaction with IEP. The complexes with the nicked WT DNA substrate showed the same two bend angles (74 ± 2°, 50% and 89 ± 2°, 50%) as those with the non-nicked

WT substrate, while complexes formed with the nicked +4/+15 DNA substrate still showed a wide distribution of bend angles (Figure 3.9). These experiments demonstrate that the bottom-strand nick is not by itself sufficient to cause bending and together with the results for the ΔEn LtrA mutant with the wild-type DNA substrate (Figure 3.8) indicate that bending requires specific interactions between the IEP and 3'- exon.

### 3.3.4 AFM of RNP-DNA complexes after initiation of reverse transcription

The addition of dNTPs during the retrohoming reaction allows LtrA to initiate TPRT using the cleaved bottom strand as a primer. Adding just three dNTPs (dATP, dTTP and dGTP) and omitting dCTP arrests reverse transcription opposite the first C-residue in the template RNA (17 nts), which results in uniform complexes held at the same stage of the reaction. Complexes formed by incubating wild-type RNPs with the WT DNA substrate in the presence of the three dNTPs for 5 and 30 min, respectively, were imaged by AFM (Figure 3.10). After 5 min, the number of complexes with the two major bend angles was roughly equal (76 ± 3°, 48% and 90 ± 2°, 52%), whereas after 30 min, the proportion of molecules with the greater bend angle increased dramatically (73 ± 2°, 16% and 90 ± 4°, 84%). By contrast RNPs reconstituted with the $DD^-$ LtrA mutant, which have the mutation YADD→YAAA at the RT active site (Matsuura et al. 1997), with WT DNA after 30 min incubation under the same conditions, were predominantly found bent at the smaller angle (73 ± 2°, 46 %) with only a small number of complexes bent at the greater angle (90 ± 2°, 11%) (Figure 3.11). Together, these results suggest that the greater bend angle is characteristic of RNPs carrying out reverse transcription. The results for $DD^-$ RNPs may reflect that the integrity of the RT active site is required to

achieve the greater bend angle, but we cannot exclude the possibility that the mutations affect global protein structure.

### 3.3 DISCUSSION

Imaging Ll.LtrB RNP-DNA substrate complexes using AFM demonstrates that the DNA target site is bent during the retrohoming pathway. The length of the DNA in the complex as determined by non-deconvoluted AFM measurements (85 ± 9 nm) is in good agreement with the calculated length (77.2 nm) and close to that determined by AFM measurements for free DNA substrate (81 ± 11 nm), indicating that the DNA is not wrapped around the RNPs to an appreciable extent.

The Ll.LtrB RNPs bound to DNA appear roughly circular with an average diameter of ~10 nm and a height of ~1 nm. The diameter of the Ll.LtrB RNPs (0.45 MDa), which consist of a 0.9-kb RNA and two molecules of the 80 kDa LtrA protein (Saldanha et al. 1999), is in reasonable agreement with expectations for an RNP complex based on the *E. coli* small and large ribosomal subunits. The dimensions of the *E. coli* small ribosomal subunit (0.85 MDa) determined by cryoelectron microscopy or X-ray crystallography are 10 x 15 x 6 nm, and those for the large ribosomal subunit (1.7 MDa) are 24 x 24 x 15 nm (Lata et al. 1996; Ban et al. 1998; Verschoor et al. 1998; Tocilj et al. 1999). The relatively low height measured for Ll.LtrB RNPs may reflect compaction during air drying/imaging or compliance differences in tapping mode between the mica surface and the RNP. *E. coli* small ribosomal subunits appear to be flattened by about 50% by AFM imaging in air versus AFM imaging in solution (Matsuura et al. 2006).

Ll.LtrB RNPs of ~10 nm diameter could potentially contact 30-bp of unbent DNA or >30-bp if the DNA is bent, in agreement with known distal contact points at positions -23 and +10 (Singh and Lambowitz 2001). DNase I-footprinting, which may somewhat overestimate the size of the binding site, showed protection over 40-45 bp (positions -25 to +19 on the top strand and -28 to +16 on the bottom strand) (Singh and Lambowitz 2001). The AFM images emphasize the striking compaction of the 902-nt intron RNA in RNPs, which appear as a small dot on the 234-bp double-stranded DNA substrate.

The computational analysis showing that both retrohoming and retrotransposition sites have regions of higher predicted bendability than do random *E. coli* DNA sequences suggests that bendability is a significant parameter in the use of these sites. Bendability does not distinguish whether the RNPs preferentially recognize DNA target sites that were already bent or whether the flexibility of the target site increases the efficiency of the reaction, by decreasing the energy required to achieve the bend. AFM comparison of bend angles at the same position in the wild-type and -30/+15 mutant DNA substrates in absence of RNPs showed higher bend angles for the wild-type DNA (Figure 3.6), consistent with the possibility that the wild-type target site contains a higher proportion of molecules pre-bent at the correct positions. Assuming that the DNA on the surface represents a snapshot of the state of DNA bending in solution, these AFM results agree with recent simulations of the in-plane bending of short semiflexible polymers (Lattananzi 2004). While polymers with a ratio ("t") of contour length to persistence length of about two ($t \approx 2$) can only probe bend angles up to 90° with significant probability, more flexible polymers that have reduced persistence lengths probe larger

bend angles. Given that the persistence length of DNA is generally 40-45 nm or about half the length of these DNA substrate molecules, the distribution of bend angles in the -30/+15 DNA show the predicted behavior for t = 2, while the wild-type DNA shows the behavior expected for t > 2 or much reduced stiffness (Wang et al. 1997).

Both the retrohoming and retrotransposition sites contain only a small number of nucleotide residues whose bases are recognized directly by the IEP, and with the exception of T+5, these are only moderately conserved in different sites (*e.g.*, T-23, 41%; G-21, 79%; and A-20, 61% in the 66 retrohoming sites, and T-23, 15%; G-21, 48%; and A-20, 33% in the retrotransposition sites) (Zhong et al. 2003; Coros et al. 2005). The region of the retrohoming sites recognized by base pairing of the intron RNA also has variable sequences, due to the method by which the sites were selected (*i.e.*, with an Ll.LtrB intron having randomized EBS2, EBS1, and δ sequences). Thus, it seems likely that bendability itself is a factor in the selection of Ll.LtrB intron target sites, beyond the nucleotide residues recognized directly by Ll.LtrB RNPs.

Figure 3.12 summarizes these AFM results in the form of a model showing the degree of bending at different stages of the reaction. First, initial DNA target site recognition and reverse splicing do not by themselves lead to DNA bending. A mutant DNA substrate with alterations between 3'-exon positions +4 and +15 (+4/+15), and RNPs reconstituted with mutant LtrA that lacks the En domain carry out these initial steps with relatively high efficiency, but are unable to carry out bottom-strand cleavage. In both cases, AFM imaging shows complexes with RNPs bound at the correct position, but with a wide distribution of DNA bend angles, similar to naked DNA.

Second, DNA bending likely results from the force generated by RNPs binding simultaneously to both the 5'- and 3'-exon regions of the DNA target site. Previous studies showed that all mutations or chemical modifications in the distal 5'-exon or IBS/δ region of the DNA target site that affect LtrA interactions or base pairing of the intron RNA inhibit not only reverse splicing, but also bottom-strand cleavage (Mohr et al. 2000; Singh and Lambowitz 2001). Although we cannot exclude that some 5'-exon interactions are weakened or lost at later stages, a likely interpretation is that most if not all of the initial RNP interactions with the distal 5'-exon are required along with 3'-exon interactions to position LtrA En domain for bottom-strand cleavage. Our experiments with nicked DNA substrates confirm that bending results from RNP interactions with the 3'-exon and not from bottom-strand cleavage itself (Figure 3.8). As indicated previously, three-dimensional modeling predicted that LtrA could not interact simultaneously with its target sequences in the distal 5'-exon and the 3'-exon unless the DNA is bent strongly in the complex (Blocker et al. 2005).

The two progressively greater bend angles seen in time-course experiments presumably reflect sequential LtrA interactions that are known to occur with the 3'-exon. These are recognition of T+5 and 3'-exon positions for bottom-strand cleavage, binding of the scissile phosphate between 3'-exon positions +9 and +10 at the En active site, and finally, repositioning of the 3'-end of the cleaved bottom strand at the RT active site. The latter is expected to require a significant conformational change in the complex. We find that the proportion of complexes containing the greater bend angle increases after initiation of reverse transcription by wild-type Ll.LtrB RNPs, but decreases for

complexes formed with RNPs reconstituted with the DD$^-$ LtrA mutant, which have a mutation in the RT active site (Figure 3.9). X-ray crystal structures of retroviral RTs with bound template-primer substrates typically show a 40° bend in the template strand near the polymerase active site (Ding et al. 1998; Cote et al. 2000; Sarafianos et al. 2001) and even greater bend angles are seen for other polymerases (*E. coli* RNA polymerase, 90° and human DNA polymerase β, 90°) (Maitra et al. 2002). The DNA bend angle observed here for reverse transcribing RNPs may reflect not only an analogous bend at the RT active site, but also that the RNPs maintain contacts involved in DNA target site recognition during the initial stages of reverse transcription. At present, one can only imagine the acrobatics required to progressively unwind the highly folded intron RNA through the RT active site, a process that presumably requires disruption of the strong LtrA interactions that stabilized the folded intron RNA structure for RNA splicing and reverse splicing, as well as LtrA and intron RNA contacts with the DNA target site.

## 3.4 METHODS

### 3.4.1 DNA substrates

DNA substrates used for biochemical assays and AFM experiments were generated by PCR of pLHS, which contains a 70-nt sequence of ligated *ltrB* exons 1 and 2, extending from 35-nt upstream to 35-nt downstream of the intron-insertion site (positions -35 to +35), cloned between the BamHI and EcoRI sites of pBSKS+ (Agilent Technologies, Santa Clara, CA) (Matsuura et al. 1997).

To generate the wild-type DNA substrate, the PCR was carried out with primers SK+113 (5'-CCCTCACTAAAGGGAACAAAAAGC) and KS-121 (5'-CGACGGCCAGTGAATTGTAATAGC). The resulting 234-bp double-stranded DNA

contains the 45-bp E1-E2 target sequence flanked by 5'- and 3'-sequences of 91 and 98 bp, respectively. Target site L (240 bp) contains the 45-bp Ll.LtrB intron target sequence at one end of the DNA oligonucleotide (positions 6-50) and was generated by PCR of pLHS with primers KS-35 (5'-CTTGCAACCCACGTCGATCGTG) and SK+205 (5'-CTTCCGGCTCGTATGTTGTGTGG). Mutant DNA substrates (234 bp) contain unfavorable non-wild-type nucleotide residues at target-site positions -30 to -13, -12 to +3, +4 to +15, and -30 to +15, respectively. The non-wild-type residues are those occurring at lowest frequency at each position in the set of retrohoming sites in the *E. coli* genome obtained with an Ll.LtrB intron with randomized target-site recognition sequences (Zhong et al. 2003). These mutant target DNAs were generated via two PCR steps. In the first step, two separate PCR reactions were done with outside primers SK+113 or KS-121 and overlapping internal primers, containing the DNA target site mutations. The internal primers were: SK-30/-13 (5'-GTAAGATAATGGTCCTGACACATCCATAACCATATC)
and KS-30/-13 (5'- TGTGTCAGGACCATTATCTTACTGCAATTATCCACTAG);
SK -12/+3 (5'-TCGTGACTGAAGGTAAAAAATGATCATTTTTAATTCTACG) and
KS -12/+3 (5'-CATTTTTTACCTTCAGTCACGATCGACGTGGGTTG);
SK +4/+15 (5'- CGGCAATATATGTCTACGAATCTTTATACTGG) and KS +4/+15
(5'- CGTAGACATATATTGCCGATGGTTATGGATGTGTTCACGA); and SK -30/+15
(5' -TTTTTACCTTCAGCAGGACCATTATCTTACTGCAAGGATCCACTAG) and KS -30/+15 (5'-GGTCCTGCTGAAGGTAAAAAATGCGGCAATATATGTCTACGAATCTTTATAC TGGG). The products of the first PCR were gel-purified, mixed, and used in a second PCR with the outside primers to yield 234-bp DNA substrates containing the desired modifications.

To generate wild-type and mutant +4/+15 DNA substrates containing a nick at the bottom-strand cleavage site, the top strand was amplified by unidirectional PCR of pLHS or the gel-purified +4/+15 DNA substrate, respectively, with primer KS-121 (see above). After extraction with phenol-CIA, the unidirectional PCR products were annealed with a ten-fold molar excess of complementary gel-purified bottom-strand oligonucleotides (129 and 105 nts) by heating to 90°C for 5 min and then slowly cooling to room temperature. The resulting 234-bp double-stranded DNAs with a nick at the bottom-strand cleavage site were purified in a 1% agarose gel.

**3.4.2 Reconstitution of RNPs**

Ll.LtrB RNPs were reconstituted with purified LtrA protein and *in vitro* synthesized Ll.LtrB-ΔORF lariat RNA. Wild-type and mutant LtrA proteins were expressed in *E. coli* BL21(DE3), by using the intein-based expression plasmid pImp-1P, and purified as described in Saldanha et al. (1999). The LtrA mutant $DD^-$ has the conserved amino acid residues YADD in the RT active site changed to YAAA (Matsuura et al. 1997), and mutants ΔEn and ΔD/En have C-terminal truncations at amino acid residues 547 and 490, respectively, which delete the En and D + En domains, respectively (San Filippo and Lambowitz 2002). Protein concentrations were determined by Bradford assay (Bradford 1976), using LtrA protein whose concentration had been determined by using $A_{280}$ as a standard. The protein preparations were >98% pure, as judged by Coomassie blue-stained SDS-polyacrylamide gels.

Ll.LtrB-ΔORF lariat RNA was generated by self-splicing of a 971-nt Ll.LtrB RNA (902-nt Ll.LtrB intron flanked by 5'- and 3'-exon sequences of 32 and 37 nt, respectively) transcribed from BamHI-linearized pJNΔORF with phage T7 RNA

polymerase, using a MEGAScript kit (Ambion) (Noah and Lambowitz 2003). Self-splicing was carried out with 2 μM Ll.LtrB-ΔORF RNA in 1 ml of 0.5 M $NH_4Cl$, 50 mM $MgCl_2$, 40 mM Tris-HCl, pH 7.5 at 30°C. After renaturing the RNA by heating to 50°C for 1 min and then slowly cooling to 30°C in 0.5 M $NH_4Cl$, 5 mM $MgCl_2$, 40 mM Tris-HCl, pH 7.5, self-splicing was initiated by adding the remaining $MgCl_2$. The reaction was incubated for 2 h at 37°C and terminated by extraction with phenol-CIA.

To reconstitute RNPs, the LtrA protein was incubated with the *in vitro* self-spliced lariat RNA in 15 ml of reaction medium containing 0.5 M $NH_4Cl$, 5 mM $MgCl_2$, 40 mM HEPES, pH 7.5. The RNA (3 nmoles in 1 ml of reaction medium) was renatured by heating to 50°C for 1 min and then slowly cooling to 30°C. After adding LtrA protein (9 nmoles), the sample was diluted to 15 ml with reaction medium and incubated for 2 h at 30°C. RNPs were then pelleted by centrifugation in a Beckman Ti50.2 rotor (Beckman-Coulter; 40,000 rpm, ≥ 8 h at 4°C). The RNP pellet was dissolved in 300 μl of 10 mM KCl, 10 mM $MgCl_2$, 40 mM HEPES, pH 7.5, 50% glycerol, aliquoted, and stored at ≤ -20°C. A typical RNP preparation contained 88% lariat RNA, as judged by electrophoresis in a denaturing 4% polyacrylamide gel.

### 3.4.3 DNA integration assay

Reactions were carried out by incubating reconstituted RNPs (6 pmol) with gel-purified DNA substrates (6 pmol) in 100 μl of 10 mM KCl, 10 mM $MgCl_2$, 40 mM HEPES, pH 7.5 (Saldanha et al. 1999). The reactions were initiated by adding RNPs, incubated for 5 or 30 min at 37°C, and terminated by phenol-CIA extraction and ethanol

precipitation. The products were analyzed in a denaturing 4% polyacrylamide gel, which was scanned with a PhosphorImager.

### 3.4.4 Atomic force microscopy

For AFM, 6 pmol of RNPs (based on RNA $A_{260}$) were incubated with 6 pmol of target DNA in 100 μl of 10 mM KCl, 10 mM $MgCl_2$, 40 mM HEPES, pH 7.5 for 5 or 30 min at 37°C, then placed on ice. Samples were adsorbed onto aminosilane-modified mica surfaces using a slight modification of the procedure described in by (Balagurumoorthy et al. 2002). Freshly cleaved mica was exposed to the vapor of 3-aminopropyltriethoxysilane (98%; Sigma-Aldrich) for 1 h, then baked for 1 h at 100°C. After diluting RNP-DNA complexes to 650 pM in 10 mM KCl, 10 mM $MgCl_2$, 40 mM HEPES, pH 7.5, a 200 μl aliquot was applied to the treated mica slide, incubated on ice for 20 min, washed once with 500 μl buffer, twice with 500 μl water, and air dried. AFM imaging was done in tapping mode using a MFP-3D scanning probe microscope (Asylum Research, Santa Barbara, CA). Super Sharp Silicon 225 μm tips (SSS-NCLR or SSC-NCL, Nanoworld, Neuchatel, Switzerland) with a 190 kHz frequency were used for imaging. We discerned no difference in resolution with the different tips.

DNA bend angles were measured using Scion Image software (Scion, Frederick, MD). Only complexes that were completely visible and could be traced unambiguously were used for analysis. Selected images of single complexes were enlarged and two tangential lines were drawn along the center of the protruding DNA arms on both sides of the globular RNP, to give the measured bend angle. The angles were defined by manually determining the path of the DNA segments exiting the RNP within about 15 nm on each

side (Janicijevic et al. 2003). For naked DNA, a circular mask about the size of the RNP was placed over the center of the DNA, and the DNA bend angle at this position was determined as described above The angle apex was chosen as the point were the DNA path trajectories intersect. In each case, the DNA bend angle was calculated from measurements of 100 molecules, and the values are expressed as the mean ± the standard deviation. By convention, the bend angle is defined as deviation from straight.

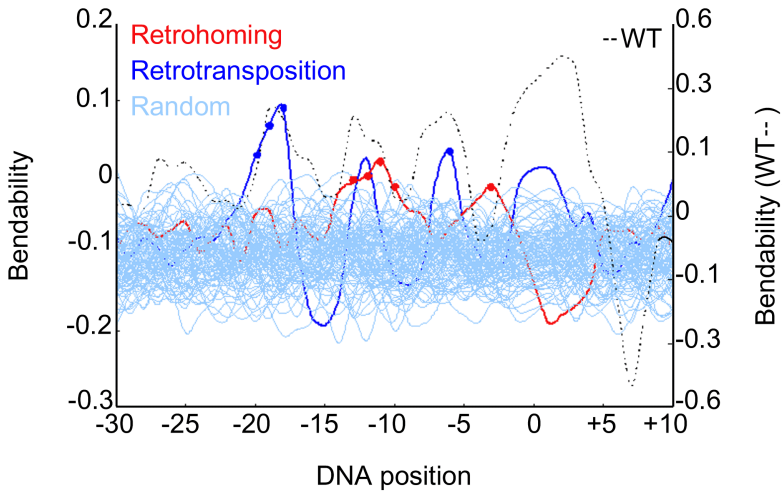


Figure 3.1: DNA bendability profiles of the wild-type Ll.LtrB intron DNA target site compared to non-wild-type retrohoming sites, retrotransposition sites, and random DNA

The profiles show average bendability versus nucleotide position calculated for sets of 66 previously identified Ll.LtrB intron retrohoming sites (red line) and 33 Ll.LtrB intron retrotransposition sites (blue line). The bendability profiles were compared to those for 100 control sets consisting of the same number of randomly selected *E. coli* DNA sequences (6600 and 3300 individual sequences, respectively). The profiles for the 100 retrohoming control sets are shown as light blue lines. The profiles for the 100 retrotransposition control sets are similar to those for the retrohoming control sets and are not shown in order to simplify the figure. Bendability profiles were calculated over a 6-bp sliding window for positions -30 to +10 from the intron-insertion site by using an algorithm that employs a probabilistic model based on DNase I-susceptibility of different trinucleotides. Dots mark the nucleotide position at the 5'-end of six-bp segments that are significantly more bendable in the retrohoming or retrotransposition sites than in their respective control sets, as determined by a statistical test for the difference between two sample means using a 5% significance level. The calculated bendability profile for the wild-type Ll.LtrB intron DNA target site (dashed line) is shown for comparison (note difference in scale on the right) (Noah et al. 2006).

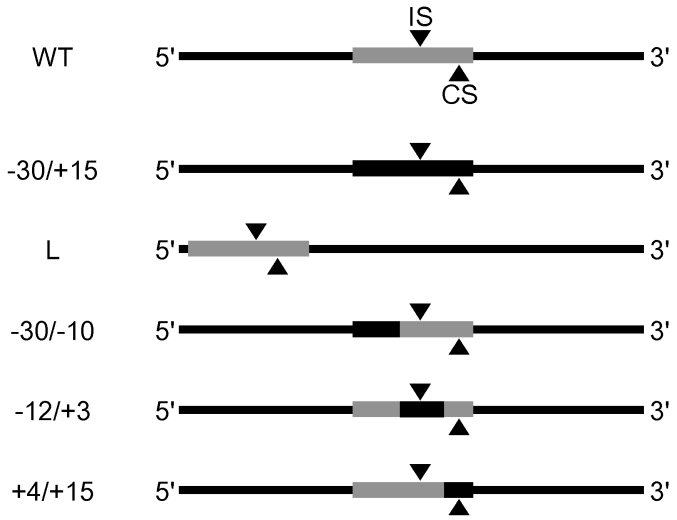


Figure 3.2: Schematic of DNA substrates used in AFM experiments

Substrates are ~240-bp double-stranded DNA molecules generated by PCR. The wild-type Ll.LtrB DNA target site (positions -30 to +15 relative to the intron insertion site) is shown as a gray rectangle, with black shading showing regions changed in mutant DNA substrates. In DNA substrate L, the wild-type Ll.LtrB intron target site is moved near the 5'-end of the DNA molecule. Triangles above and below denote the intron insertion site (IS) and bottom-strand cleavage site (CS), respectively.

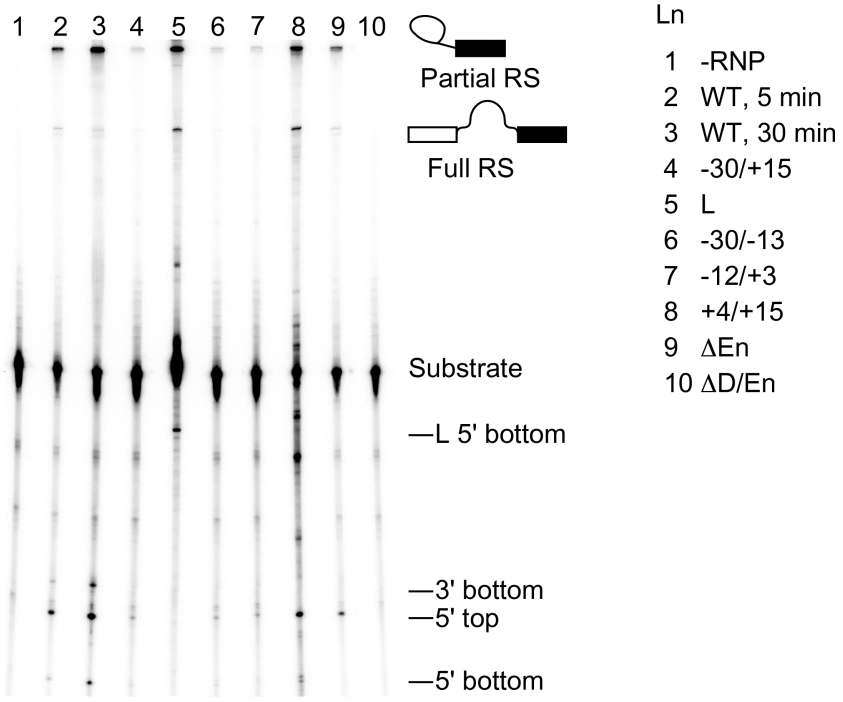


Figure 3.3: Reverse splicing and bottom-strand cleavage assay with DNA substrates used for AFM

Wild-type or mutant Ll.LtrB RNPs were incubated with internally $^{32}$P-labeled DNA substrates (Figure 3.2) for 5 min (lane 2) or 30 min (lanes 3-10) at 37°C and the products were analyzed in a denaturing 4% polyacrylamide gel. Lane 1 shows the WT DNA target substrate incubated in reaction medium for 30 min at 37°C without RNPs. The substrates used in the reactions for lanes 2-10 are indicated on the far left. Positions of products resulting from partial and complete reverse splicing and bottom-strand cleavage are indicated to the left of the gel (Noah et al. 2006).

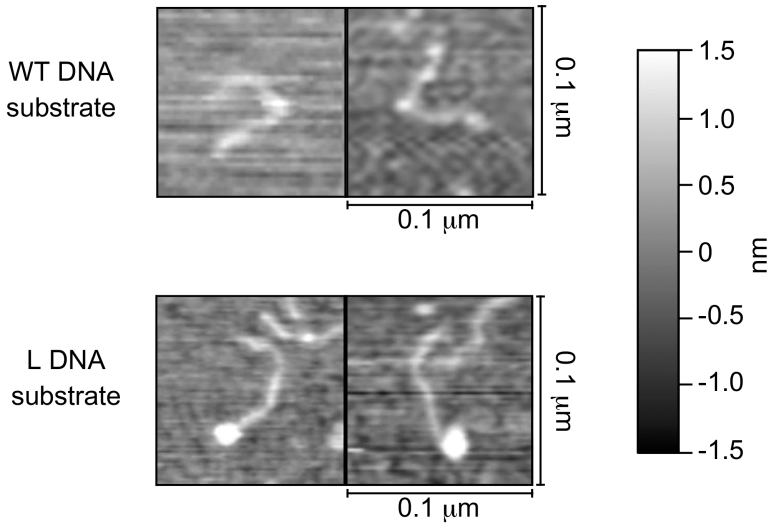


Figure 3.4: Representative AFM images of RNP-DNA complexes with WT and L DNA substrates

AFM images of wild-type Ll.LtrB RNPs bound to DNA substrates in which the Ll.LtrB intron target sequence is at the center (WT) or one end (L) of the DNA molecule, respectively (see Figure 3.2), taken in tapping mode over a single $\mu m^2$. The scale to the right of the images denotes the sample height in nm and is consistent throughout all AFM images shown in proceeding figures.

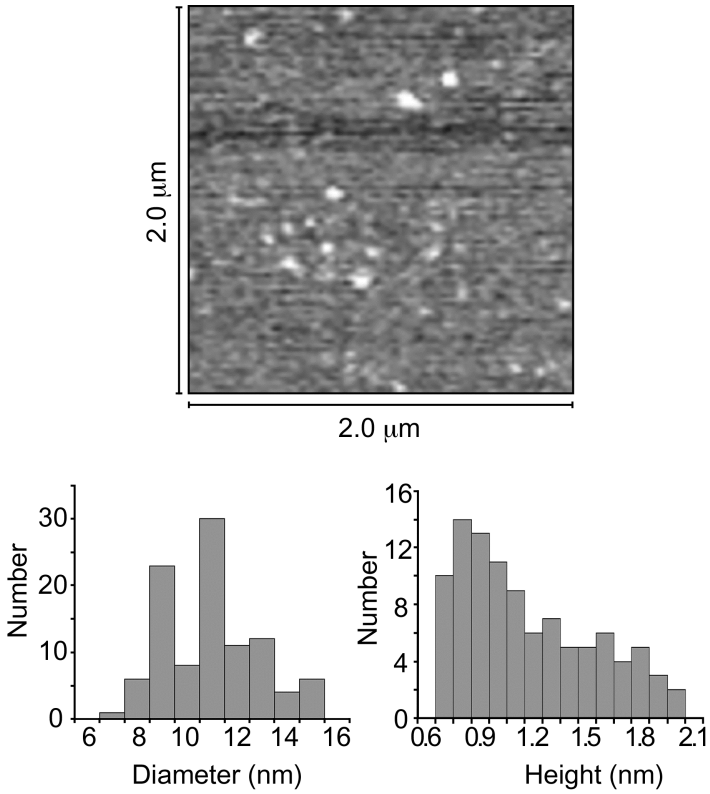


Figure 3.5: AFM image of free Ll.LtrB RNPs

The figure shows a representative field of RNPs (above) imaged as in Figure 4, after incubation in 10 mM KCl, 10 mM $MgCl_2$, 40 mM HEPES, pH 7.5 for 5 min at 37°C. The bar graphs (below) show diameter and height measurements for 100 imaged RNPs.

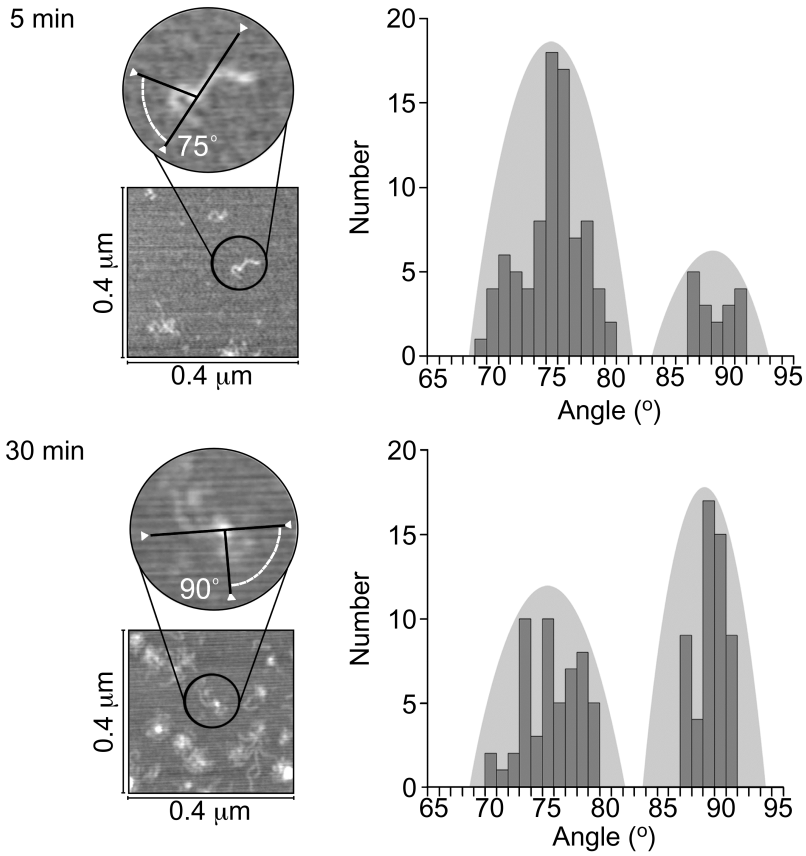


Figure 3.6: AFM imaging of complexes formed with wild-type Ll.LtrB RNPs and DNA substrate for different times

WT DNA substrate was incubated with Ll.LtrB RNPs for 5 min at 37°C, and 30 min at 37°C and imaged by AFM. Partial field views (0.4 $\mu m^2$) are shown for each. The expanded views of single complexes for each time point highlights a RNP-DNA complex with a representative DNA bend angle at that time point, 78° at 5 min and 90° at 30 min. The histograms on the right of each image show the distribution of DNA bend angles in 100 complexes for each time point. Gray shaded bell-curves show the bimodal distribution of DNA bend angles.

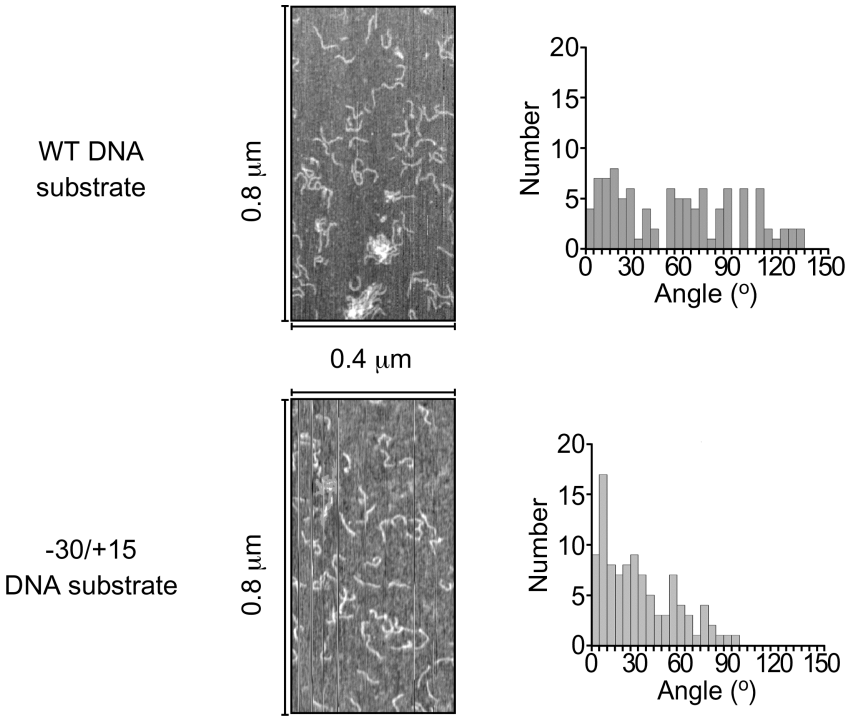


Figure 3.7: AFM images of naked DNA substrates

WT and -30/+15 DNA substrates in the absence of RNPs were imaged by AFM. Partial field views (0.8 μm x 0.4 μm) are shown for each. The histograms to the right of each image show the distribution of DNA bend angles of 100 molecules. The range of DNA bend angles of the WT DNA substrate includes larger angles than those observed with the -30/+15 DNA substrate, perhaps reflecting the greater inherent bendability of the wild-type Ll.LtrB intron target site.

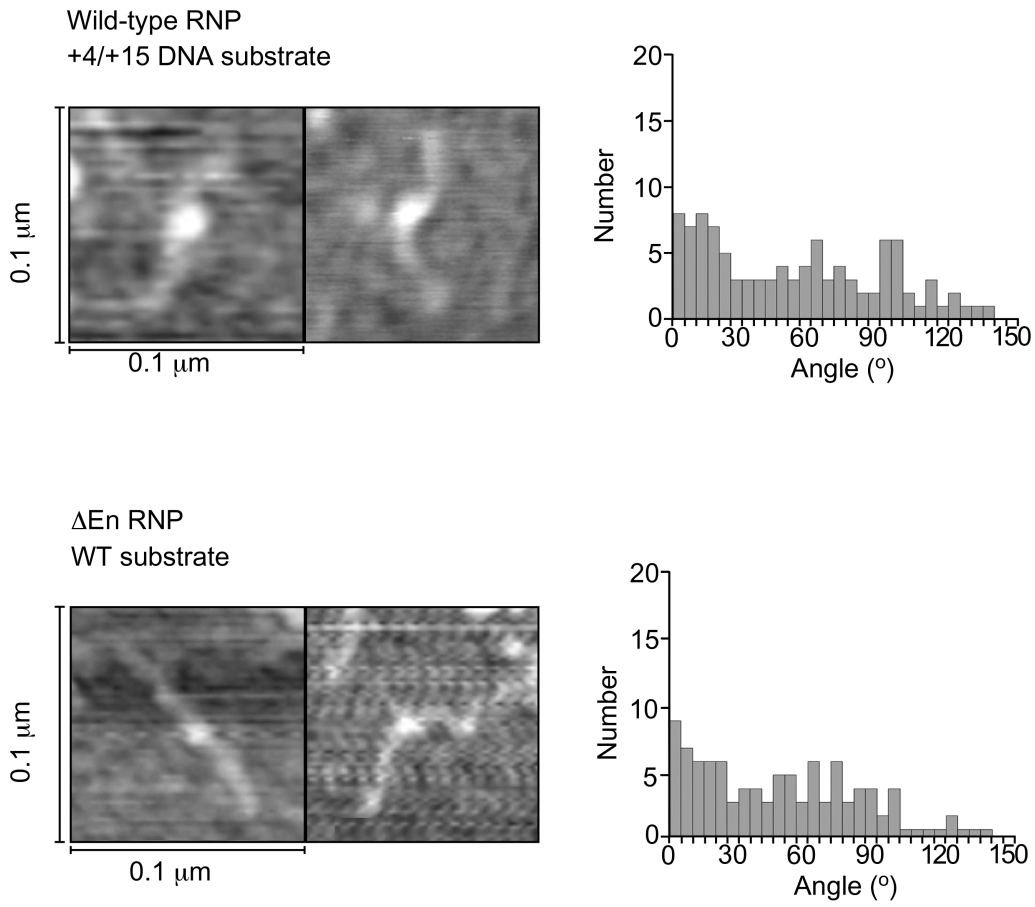


Figure 3.8: AFM images of RNP-DNA complexes with disrupted 3'-exon interactions

To determine the effect of disrupting RNP interactions with 3'-exon of the DNA target site, two sets of DNA-RNP complexes were used. First, wild-type Ll.LtrB RNPs were incubated with the +4/+15 DNA substrate, which has disfavored nucleotide-residues at 3'-exon positions +4 to +15, and imaged by AFM. Next, WT DNA substrate was incubated with ΔEn RNPs, which are reconstituted with a mutant LtrA protein lacking the C-terminal En domain, and were imaged by AFM. The histograms to the right of the images show the distribution of DNA bend angles in 100 complexes for each condition. DNA bend angles in both sets of complexes with disrupted 3'-exon interactions are distributed similarly to that of naked (unbound) DNA (Figure 3.7). All incubations were for 30 min at 37°C.

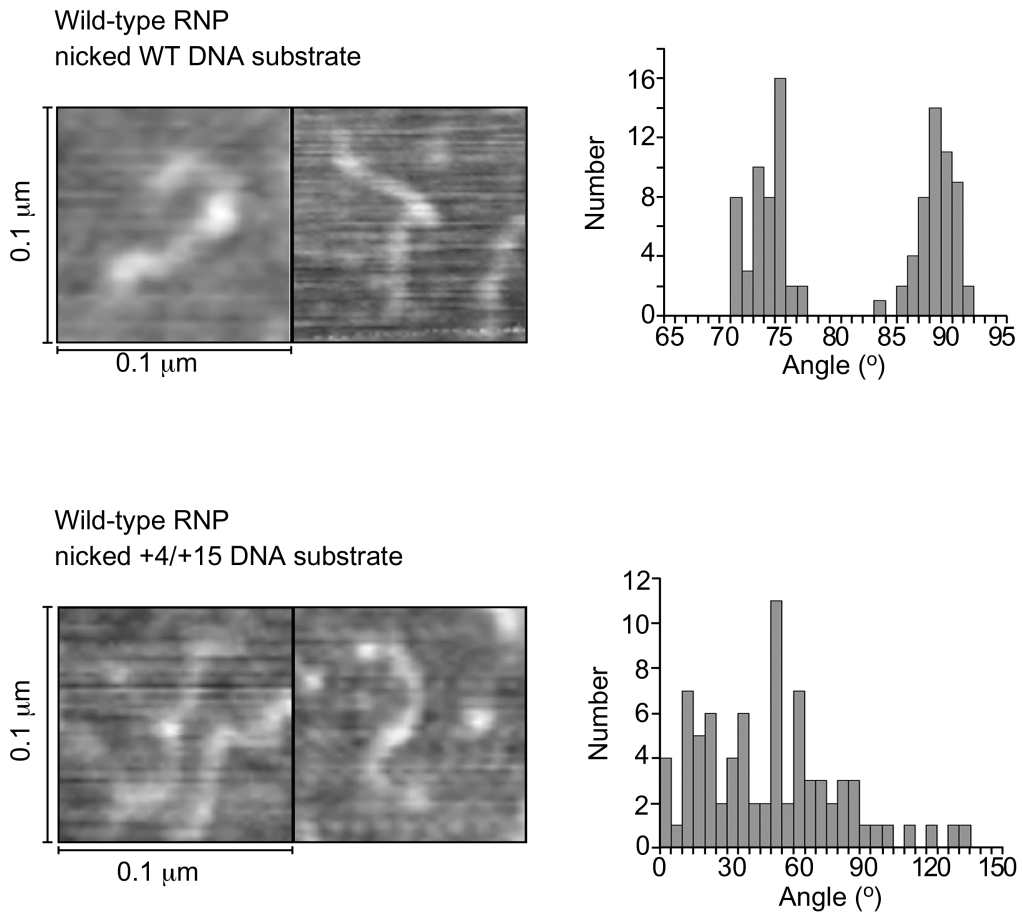


Figure 3.9: AFM images of RNP-DNA complexes with nicked DNA substrates

Wild-type Ll.LtrB RNPs were incubated with WT and +4/+15 mutant DNA substrates, both containing a pre-positioned nick at the bottom-strand cleavage site, for 30 min at 37°C and imaged with AFM. The histograms to the right of the images show the distribution of DNA bend angles in 100 complexes for each nicked DNA substrate. The distribution of DNA bend angles in complexes with the nicked WT DNA substrate mirrors that of complexes with non-nicked WT DNA substrate with two predominate clusters of angles at approximately 75° and 90° (Figure 3.6, 30 min). The DNA bend angles in complexes with the nicked +4/+15 DNA substrate are broadly distributed mirroring the distribution of DNA bend angles in complexes with the non-nicked +4/+15 DNA substrate (Figure 3.8).

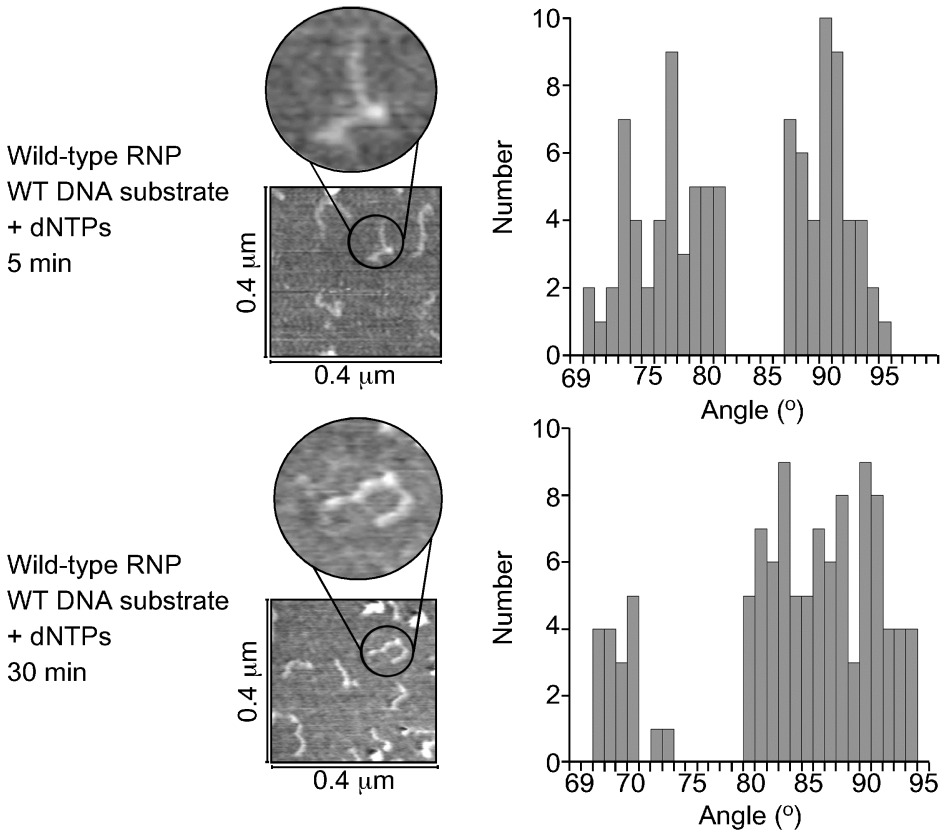


Figure 3.10: AFM images of RNP-DNA complexes after initiation of reverse transcription

Wild-type Ll.LtrB RNPs were incubated with WT DNA substrate at 37°C in the presence of 2 μM each dATP, dTTP, and dCTP for 5 and 30 min, respectively, and imaged by AFM. The highlighted molecules illustrate the predominant bend angle at each time, 78° at 5 min and 90° at 30 min. With wild-type RNPs the presence of dNTPs skews the distribution of DNA bend angles toward the larger angle (90°) at 5 min compared to complexes imaged in the absence of dNTPs (Figure 3.6). The distribution is even more heavily skewed toward the 90° bend angle with wild-type RNPs in the presence of dNTPs incubated for 30 min compared to complexes imaged without dNTPs (Figure 3.6).

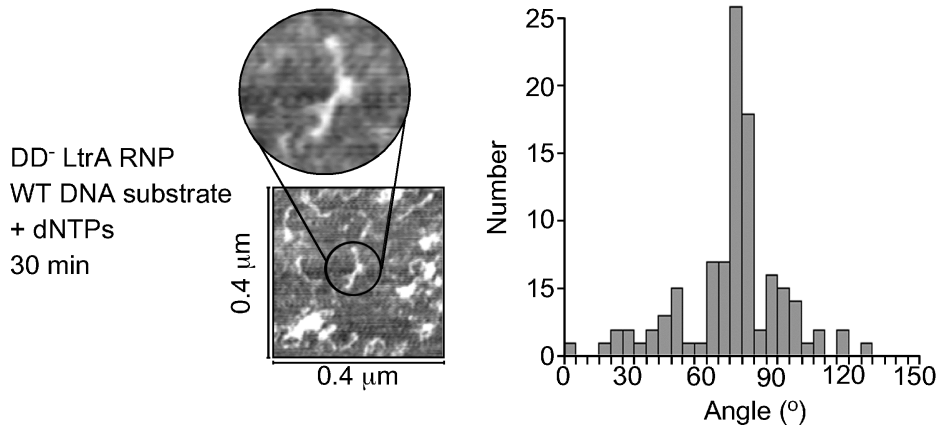


Figure 3.11: AFM images of RNP-DNA complexes with RT-deficient LtrA protein

Ll.LtrB RNPs reconstituted with the DD- LtrA mutant were incubated with WT DNA substrate at 37°C in the presence of 2 μM each dATP, dTTP, and dCTP for 30 min and imaged by AFM. These complexes do not show the same DNA bend angle distribution pattern as those with wild-type RNPs. In a large population of these complexes, the DNA is bent at ~75°, but relatively few are bent at the larger 90° angle.

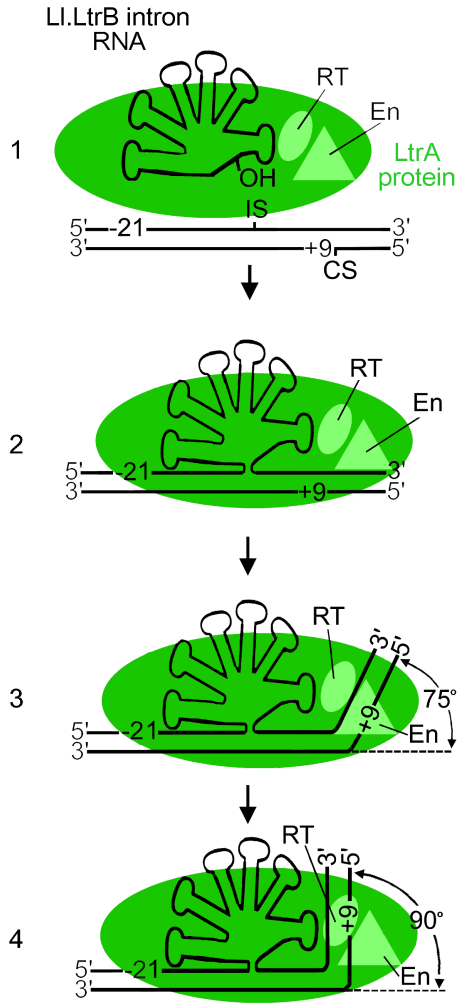


Figure 3.12: DNA bending during Ll.LtrB intron retrohoming

(1) RNPs bind DNA non-specifically and search for target sites. (2) Initial DNA target site recognition involves IEP interactions with the distal 5'-exon region and base pairing of the intron RNA's EBS2, EBS1, and δ sequences to the IBS2, IBS1, and δ' sequences between positions -12 to +3 of the DNA target site. These initial DNA target site interactions and reverse splicing of the intron RNA into the intron-insertion site (IS) in the DNA top strand do not lead to detectable DNA bending. (3) Subsequent IEP interactions with the 3'-exon lead to an initial DNA bend angle of 75° and bring the scissile phosphate at the bottom-strand cleavage site (CS) to the En active site. (4) Shift of the 3'-end of the cleaved bottom strand from the En to RT active site for initiation of reverse transcription leads to a greater bend angle of 90°.

# Chapter 4: Gene targeting in *Bacillus subtilis* using reprogrammed Ll.LtrB introns

Reprogrammed Ll.LtrB group II introns ("targetrons") have been used for site-specific insertion and gene disruption in a variety of organisms (Karberg et al. 2001; Perutka et al. 2004; Yao et al. 2006; Yao and Lambowitz 2007; Mastroianni et al. 2008). Here, I demonstrated the suitability of targetrons for gene targeting in the Gram-positive bacterium, *B. subtilis* by the insertion of the intron into two genes, *yhcS* and *ywpE*. These two genes where chosen because each has a corresponding homolog in pathogenic Gram-positive species such as the closely related *Bacillus anthracis* (Comfort and Clubb 2004). Recently, interest has grown in the *B. anthracis* homologs of *yhcS* and *ywpE*, *srtA* and *srtB*, respectively, due to their role in bacterial virulence (Zink and Burns 2005; Maresso and Schneewind 2008).

These genes encode sortase enzymes, which are transpeptidases responsible for anchoring proteins to the bacterial cell wall. The *yhcS*/*srtA* gene product, Sortase A (SrtA), is a general house-keeping enzyme responsible for attaching a variety of protein types to the cell wall (Comfort and Clubb 2004; Dramsi et al. 2005). The *ywpE*/*srtB* gene product, Sortase B (SrtB), specifically anchors proteins involved in iron acquisition to the cell wall (Marraffini et al. 2006; Maresso et al. 2008). Certain mutations in either protein can dramatically inhibit bacterial growth in mouse macrophage cells, but sporulation and growth in BHI medium are unaffected (Zink and Burns 2005). This phenotype makes these proteins attractive drug targets because germination of *B. anthracis* spores in

alveolar macrophages is critical in the infectious progression of inhalation anthrax (Guidi-Rontani et al. 1999).

Although closely related to *B. anthracis*, *B. subtilis* is non-pathogenic. *B. subtilis* was used here as a model for the application of targetrons in *Bacillus* genus gene targeting. Both SrtA and SrtB have been reported to be non-essential for *B. subtilis* growth in culture media (Kobayashi et al. 2003).

### 4.1 INTRON MOBILITY IN *B. SUBTILIS*

To test the mobility of the reprogrammed Ll.LtrB intron in *B. subtilis*, the intron and IEP were expressed from the donor plasmid pNL9161, which was used previously for gene targeting in *Staphylococcus aureus* (Figure 4.1)(Yao et al. 2006). pNL9161 is efficiently propagated in *E. coli*, to which it confers ampicillin resistance, then purified for electro-transformation into *B. subtilis*, to which it confers erythromycin resistance. The Ll.LtrB intron with flanking 5'- and 3'-exons is expressed using a cadmium-inducible promoter. The LtrA protein is also expressed from pNL9161 downstream of the 3'-exon.

For the *B. subtilis* genomic mobility assay, expression of the intron/IEP was induced with 10 μM $CdCl_2$ at 37°C. Aliquots of the culture were then diluted and plated on LB media containing erythromycin. Targeted gene insertions were detected by colony PCR using primers that anneal to genomic DNA at positions flanking the insertion site. The insertion events were further confirmed by sequencing of the PCR amplification products.

Suitable target sites were found in both the *yhcS* and *ywpE* genes using the computer algorithm described in Perutka et al. (2004). One target site from each with the

highest predicted efficiency was chosen. The YhcS-186s targetron, which inserts at nucleotide position 186 in the sense-strand of the *yhcS* gene, was inserted into the target gene in 91 ± 5% of cells plated following cadmium induction as determined by colony PCR (Figure 4.2). Targeting by the YwpE-137a targetron, which inserts at nucleotide position 137 in the antisense-strand of the *ywpE* gene, was not detected in any colonies when plated on LB media with erythromycin, but insertion was detected when the cadmium-induced bacterial culture was used as a PCR template directly. Based on the band intensities of the PCR products from targeted and non-targeted genes in an ethidium bromide-stained agarose gel, the apparent targeting efficiency of the YwpE-137a targetron was estimated at less than 1%.

## 4.2 DISCUSSION

As demonstrated, Ll.LtrB group II intron-based targetrons can be used for gene targeting in *B. subtilis*. The two sortase enzyme genes targeted here are homologs of potentially important pathogenesis targets in *B. anthracis* and targeted gene disruption using the targetron system may provide a useful technique for the study of these genes.

Targeting of the *B. subtilis yhcS* gene is extremely efficient, with a targeting efficiency approaching 100%. As expected for a non-essential protein, the insertion events were easily detected on plated colonies after expression of the targetron. The high insertion efficiency of the YhcS-186s targetron indicates that the Ll.LtrB intron and LtrA protein function normally in *B. subtilis* when expressed from the pNL9161 donor plasmid.

The YwpE-137a targetron is active in *B. subtilis*, but with a lower apparent mobility efficiency than that of YhcS-186s. The *ywpE* gene product is only a putative SrtB protein and it has not been identified as an essential gene (Kobayashi et al. 2003). However, I find here that targeted disruption of the *ywpE* gene prevents colony growth on LB plates. The assumed function of the *ywpE* gene product, as a SrtB enzyme, is the attachment of proteins required for iron acquisition, including the secretory machinery for siderophores, to the cell wall (Marraffini et al. 2006). Cells with non-functional secretory machinery would still have access to siderophores and other secreted proteins when grown in the presence of wild-type cells in liquid culture. This may also explain the observed phenotype of *B. anthracis* SrtB mutants that are unable to grow in macrophage cells but grow normally in liquid media (Zink and Burns 2005).

## 4.3 METHODS

### 4.3.1 Recombinant plasmids and bacterial strains

The Ll.LtrB-ΔORF intron with flanking 5'- and 3'-exons was expressed from the plasmid pNL9161 downstream of a cadmium-inducible promoter (P*cad*) (Yao et al. 2006). pNL9161 is derived from the pCN37 plasmid which contains a ColE1 replicon for propagation in *E. coli* and a pT181-cop WT replicon for replication in Gram-positive species such as *B. subtilis* (Charpentier et al. 2004).

The *B. subtilis* strain ATCC6051 used in the mobility assays was provided by Dr. Richard Losick (Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA).

### 4.3.2 Targetron design

The Ll.LtrB-based targetron was retargeted to insert into *B. subtilis* genes by using a computer algorithm that identifies potential insertion sites and designs PCR primers for modifying the intron RNA to base pair to those sites (Perutka et al. 2004). Single sites in the *yhcS* and *ywpE* genes were chosen from among potential target sites identified by the algorithm. Donor plasmids were then constructed in which the intron RNA's EBS1, EBS2, and δ sequences were modified to optimize base pairing to DNA target site sequences IBS1, IBS2, and δ'. The IBS2 and IBS1 sequences in the 5'-exon of the donor plasmid were also modified to be complimentary to the retargeted EBS2 and EBS1 sequences for efficient RNA splicing (Perutka et al. 2004).

The required modifications were introduced into the donor plasmid via two-step PCR as described in Karberg et al. (2001). First, two segments of the donor plasmid were amplified by PCR. One PCR used primers p1 (5'-exon positions -25 to +18 with modifications at positions -12 to -1 for IBS1 and IBS2) and p2 (5'-CGAAATTAGAAACTTGCGTTCAGTAAAC-3'). The other PCR used primers p3 (intron positions +198 to +246 with modifications to positions +223 to +227 for EBS2) and p4 (intron positions +259 to +326 with modifications to positions +276 to +285 for EBS1 and δ). In the second step, the two PCR products from the first step were gel-purified, mixed, and amplified with primer p1 and p4 (the outer primers) to generate a 353-bp product which corresponds to the 5'-exon and 5'-end of the intron (to position +326) with modified IBS1/2, EBS2 and EBS1/δ sequences. The final PCR product was gel-purified and cloned between the BsrGI and HindIII sites of the donor plasmid.

### 4.3.3 Gene targeting in *B. subtilis*

The donor plasmids containing the retargeted introns were electroporated in to *B. subtilis* ATCC6051. Cells were grown overnight at 37°C in LB media containing erythromycin. 500 µl of the overnight culture was subcultured into 5 mL of fresh medium and grown until early log phase ($O.D._{595}$ = 0.5). Cells were then either directly plated or induced by adding 10 µM $CdCl_2$ for 90 min at 37°C prior to plating.

Targetron insertion was assayed by PCR using primers that flank the target genes and either single colonies or 1 µl of undiluted induced culture as template. For sequencing, the PCR products were run on 1% agarose gel. Bands were excised and purified using the Qiagen Gel Extraction Kit (Qiagen, Valencia, CA).

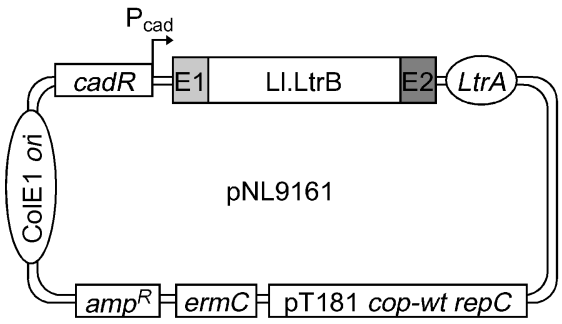


Figure 4.1: pNL9161, the targetron donor vector for *B. subtilis*

The intron donor plasmid pNL9161 uses the cadmium-inducible $P_{cad}$ promoter to express a 940-nt Ll.LtrB intron, flanked by the 5'-exon (E1) and 3'-exon (E2). The LtrA protein is expressed from a position just downstream of the 3'-exon. The ColE1 *ori* allows the plasmid to be propagated in *E. coli* for purification and transformation into *B. subtilis*. The pT181 *cop-wt repC* amplicon allows the plasmid to be replicated in Gram-positive bacteria, such as *B. subtilis* or *B. anthracis*. The plasmid carries the ampicillin resistance (*amp*$^R$) marker for *E. coli* and erythromycin resistance (*ermC*) marker for *B. subtilis* and *B. anthracis*.

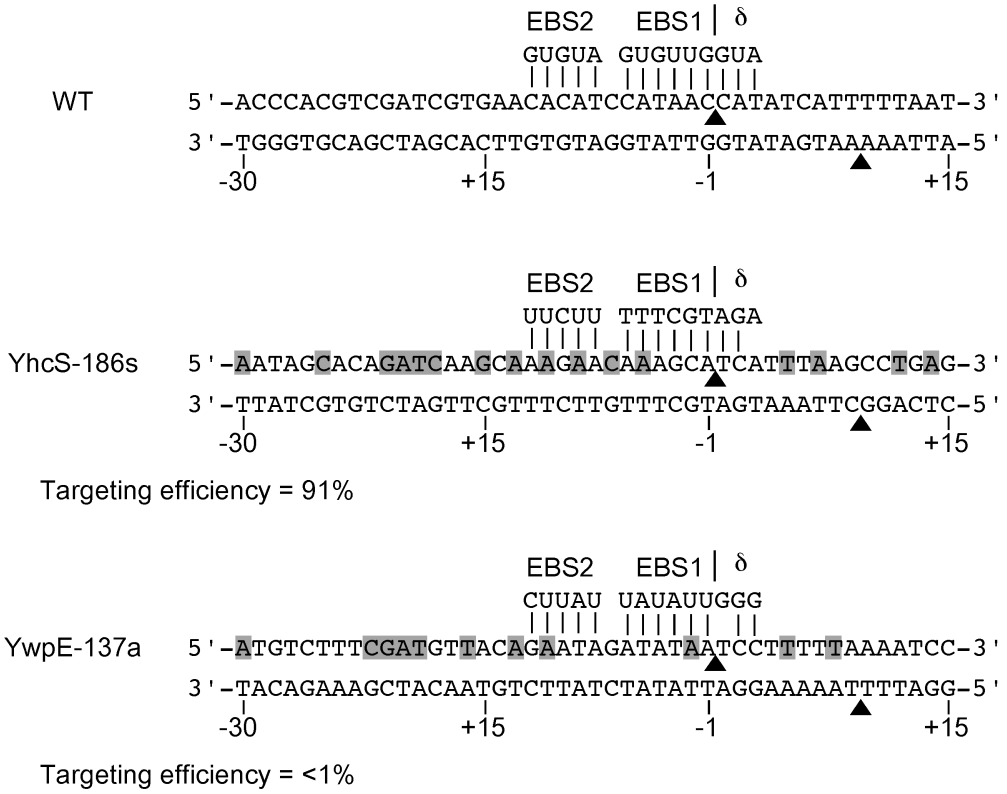


Figure 4.2: Intron base pairing in *B. subtilis* gene target sites

The wild-type Ll.LtrB intron target site with EBS-IBS/δ-δ' base-pairing interactions is shown along with the *B. subtilis yhcS* and *ywpE* target sites. Target site residues matching wild-type are shaded in gray. Black arrowheads indicate the intron-insertion site in the top strand and the predicted endonuclease cleavage site in the bottom strand. The experimental targeting efficiency of each gene target is also indicated.

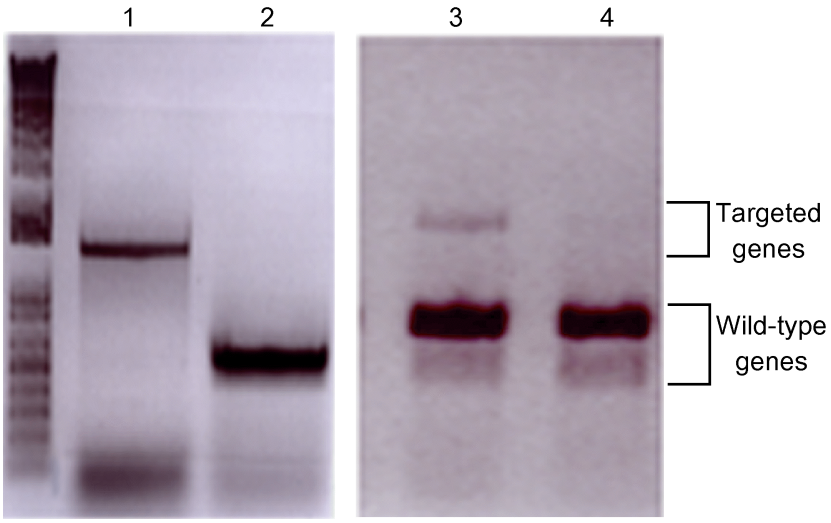


Figure 4.3: PCR detection of Ll.LtrB integration

PCR using genomic primers flanking the *yhcS* gene (lanes 1 and 2) and the *ywpE* gene (lanes 3 and 4) was used to detect Ll.LtrB integration. In lane 1, the PCR template was provided by a single colony picked from plated cells following RNP expression by 10 μM $CdCl_2$ induction at 37°C for 1 h. In lane 2, a single colony from plated cells with no pNL9161(YhcS-186s) donor plasmid was used. In lane 3, PCR template was provided by 1 μl of cell culture following RNP expression by 10 μM $CdCl_2$ induction at 37°C for 1 h. In lane 4, 1 μl of cell culture with no pNL9161(YwpE-137a) donor plasmid was used. Bands corresponding to the targeted gene (lanes 1 and 3) were gel extracted and sequenced to confirm that the LtrB intron sequence was integrated at the proper target site.

# Bibliography


Aizawa Y, Xiang Q, Lambowitz AM, Pyle AM. 2003. The pathway for DNA recognition and RNA integration by a group II intron retrotransposon. *Mol Cell* **11**(3): 795-805.

Balagurumoorthy P, Lindsay SM, Harrington RE. 2002. Atomic force microscopy reveals kinks in the p53 response element DNA. *Biophys Chem* **101-102**: 611-623.

Ban N, Freeborn B, Nissen P, Penczek P, Grassucci RA, Sweet R, Frank J, Moore PB, Steitz TA. 1998. A 9 A resolution X-ray crystallographic map of the large ribosomal subunit. *Cell* **93**(7): 1105-1115.

Bibillo A, Eickbush TH. 2002. The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J Mol Biol* **316**(3): 459-473.

Blocker FJ, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA* **11**(1): 14-28.

Bowie JH, Brinkworth CS, Dua S. 2002. Collision-induced fragmentations of the (M-H)- parent anions of underivatized peptides: an aid to structure determination and some unusual negative-ion cleavages. *Mass Spectrom Rev* **21**(2): 87-107.

Bradford MM. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**: 248-254.

Brukner I, Sanchez R, Suck D, Pongor S. 1995. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *Embo J* **14**(8): 1812-1818.

Charpentier E, Anton AI, Barry P, Alfonso B, Fang Y, Novick RP. 2004. Novel cassette-based shuttle vector system for Gram-positive bacteria. *Appl Environ Microbiol* **70**(10): 6076-6085.

Chen B, Lambowitz AM. 1997. De novo and DNA primer-mediated initiation of cDNA synthesis by the mauriceville retroplasmid reverse transcriptase involve recognition of a 3' CCA sequence. *J Mol Biol* **271**(3): 311-332.

Church GM, Sussman JL, Kim SH. 1977. Secondary structural complementarity between DNA and proteins. *Proc Natl Acad Sci U S A* **74**(4): 1458-1462.

Cole C, Barber JD, Barton GJ. 2008. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* **36**(Web Server issue): W197-201.

Comfort D, Clubb RT. 2004. A comparative genome analysis identifies distinct sorting pathways in Gram-positive bacteria. *Infect Immun* **72**(5): 2710-2722.

Coros CJ, Landthaler M, Piazza CL, Beauregard A, Esposito D, Perutka J, Lambowitz AM, Belfort M. 2005. Retrotransposition strategies of the Lactococcus lactis Ll.LtrB group II intron are dictated by host identity and cellular environment. *Mol Microbiol* **56**(2): 509-524.

Cote ML, Yohannan SJ, Georgiadis MM. 2000. Use of an N-terminal fragment from moloney murine leukemia virus reverse transcriptase to facilitate crystallization

and analysis of a pseudo-16-mer DNA molecule containing G-A mispairs. *Acta Crystallogr D Biol Crystallogr* **56 ( Pt 9)**: 1120-1131.

Cousineau B, Smith D, Lawrence-Cavanagh S, Mueller JE, Yang J, Mills D, Manias D, Dunny G, Lambowitz AM, Belfort M. 1998. Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* **94**(4): 451-462.

Cuff JA, Barton GJ. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34**(4): 508-519.

Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM. 2004. A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. *J Mol Biol* **340**(2): 211-231.

Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJ, Lambowitz AM, Zimmerly S. 2008. A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase. *Mol Cell* **30**(4): 472-485.

Dickson L, Huang HR, Liu L, Matsuura M, Lambowitz AM, Perlman PS. 2001. Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proc Natl Acad Sci U S A* **98**(23): 13207-13212.

Ding J, Das K, Hsiou Y, Sarafianos SG, Clark AD, Jr., Jacobo-Molina A, Tantillo C, Hughes SH, Arnold E. 1998. Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 A resolution. *J Mol Biol* **284**(4): 1095-1111.

Ding J, Hughes SH, Arnold E. 1997. Protein-nucleic acid interactions and DNA conformation in a complex of human immunodeficiency virus type 1 reverse transcriptase with a double-stranded DNA template-primer. *Biopolymers* **44**(2): 125-138.

Dramsi S, Trieu-Cuot P, Bierne H. 2005. Sorting sortases: a nomenclature proposal for the various sortases of Gram-positive bacteria. *Res Microbiol* **156**(3): 289-297.

el Hassan MA, Calladine CR. 1996. Propeller-twisting of base pairs and the conformational mobility of dinucleotide steps in DNA. *J Mol Biol* **259**(1): 95-103.

Eskes R, Yang J, Lambowitz AM, Perlman PS. 1997. Mobility of yeast mitochondrial group II introns: engineering a new site specificity and retrohoming via full reverse splicing. *Cell* **88**(6): 865-874.

Fedorova O, Mitros T, Pyle AM. 2003. Domains 2 and 3 interact to form critical elements of the group II intron active site. *J Mol Biol* **330**(2): 197-209.

Garvie CW, Wolberger C. 2001. Recognition of specific DNA sequences. *Mol Cell* **8**(5): 937-946.

Giro A, Bergia A, Zuccheri G, Bink HH, Pleij CW, Samori B. 2004. Single molecule studies of RNA secondary structure: AFM of TYMV viral RNA. *Microsc Res Tech* **65**(4-5): 235-245.

Gordon PM, Piccirilli JA. 2001. Metal ion coordination by the AGC triad in domain 5 contributes to group II intron catalysis. *Nat Struct Biol* **8**(10): 893-898.

Guidi-Rontani C, Weber-Levy M, Labruyere E, Mock M. 1999. Germination of Bacillus anthracis spores within alveolar macrophages. *Mol Microbiol* **31**(1): 9-17.

Guo H, Karberg M, Long M, Jones JP, 3rd, Sullenger B, Lambowitz AM. 2000. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* **289**(5478): 452-457.

Guo H, Zimmerly S, Perlman PS, Lambowitz AM. 1997. Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J* **16**(22): 6835-6848.

Huang HR, Rowe CE, Mohr S, Jiang Y, Lambowitz AM, Perlman PS. 2005. The splicing of yeast mitochondrial group I and group II introns requires a DEAD-box protein with RNA chaperone function. *Proc Natl Acad Sci U S A* **102**(1): 163-168.

Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. 2002. Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* **46**(5): 1259-1272.

Janicijevic A, Sugasawa K, Shimizu Y, Hanaoka F, Wijgers N, Djurica M, Hoeijmakers JH, Wyman C. 2003. DNA bending by the human damage recognition complex XPC-HR23B. *DNA Repair (Amst)* **2**(3): 325-336.

Jenkins BD, Kulhanek DJ, Barkan A. 1997. Nuclear mutations that block group II RNA splicing in maize chloroplasts reveal several intron classes with distinct requirements for splicing factors. *Plant Cell* **9**(3): 283-296.

Jordan SR, Pabo CO. 1988. Structure of the lambda complex at 2.5 A resolution: details of the repressor-operator interactions. *Science* **242**(4880): 893-899.

Karberg M, Guo H, Zhong J, Coon R, Perutka J, Lambowitz AM. 2001. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat Biotechnol* **19**(12): 1162-1167.

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P et al. 2003. Essential Bacillus subtilis genes. *Proc Natl Acad Sci U S A* **100**(8): 4678-4683.

Koch JL, Boulanger SC, Dib-Hajj SD, Hebbar SK, Perlman PS. 1992. Group II introns deleted for multiple substructures retain self-splicing activity. *Mol Cell Biol* **12**(5): 1950-1958.

Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. *Annu Rev Genet* **38**: 1-35.

-. 2010. Group II Introns: Mobile Ribozymes that Invade DNA. *Cold Spring Harb Perspect Biol*.

Lata KR, Agrawal RK, Penczek P, Grassucci R, Zhu J, Frank J. 1996. Three-dimensional reconstruction of the Escherichia coli 30 S ribosomal subunit in ice. *J Mol Biol* **262**(1): 43-52.

Lattananzi G, Munk., T., Frey, E. 2004. Transverse Fluctuations of grafted polymers. *Physical Reviews E* **69**: 021801.

Lenz C, Kuhn-Holsken E, Urlaub H. 2007. Detection of protein-RNA crosslinks by NanoLC-ESI-MS/MS using precursor ion scanning and multiple reaction monitoring (MRM) experiments. *J Am Soc Mass Spectrom* **18**(5): 869-881.

Luscombe NM, Laskowski RA, Thornton JM. 2001. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res* **29**(13): 2860-2874.

Luscombe NM, Thornton JM. 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* **320**(5): 991-1009.

Maitra M, Gudzelak A, Jr., Li SX, Matsumoto Y, Eckert KA, Jager J, Sweasy JB. 2002. Threonine 79 is a hinge residue that governs the fidelity of DNA polymerase beta by helping to position the DNA within the active site. *J Biol Chem* **277**(38): 35550-35560.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**(6): 793-805.

Maresso AW, Garufi G, Schneewind O. 2008. Bacillus anthracis secretes proteins that mediate heme acquisition from hemoglobin. *PLoS Pathog* **4**(8): e1000132.

Maresso AW, Schneewind O. 2008. Sortase as a target of anti-infective therapy. *Pharmacol Rev* **60**(1): 128-141.

Marraffini LA, Dedent AC, Schneewind O. 2006. Sortases and the art of anchoring proteins to the envelopes of Gram-positive bacteria. *Microbiol Mol Biol Rev* **70**(1): 192-221.

Mastroianni M, Watanabe K, White TB, Zhuang F, Vernon J, Matsuura M, Wallingford J, Lambowitz AM. 2008. Group II intron-based gene targeting reactions in eukaryotes. *PLoS One* **3**(9): e3121.

Matsuura M, Noah JW, Lambowitz AM. 2001. Mechanism of maturase-promoted group II intron splicing. *EMBO J* **20**(24): 7259-7270.

Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev* **11**(21): 2910-2924.

Matsuura T, Tanaka H, Matsumoto T, Kawai T. 2006. Atomic force microscopic observation of Escherichia coli ribosomes in solution. *Biosci Biotechnol Biochem* **70**(1): 300-302.

McCarthy JG, Williams LD, Rich A. 1990. Chemical reactivity of potassium permanganate and diethyl pyrocarbonate with B DNA: specific reactivity with short A-tracts. *Biochemistry* **29**(25): 6071-6081.

Michel F, Ferat JL. 1995. Structure and activities of group II introns. *Annu Rev Biochem* **64**: 435-461.

Michel F, Lang BF. 1985. Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. *Nature* **316**(6029): 641-643.

Mohr G, Perlman PS, Lambowitz AM. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res* **21**(22): 4991-4997.

Mohr G, Smith D, Belfort M, Lambowitz AM. 2000. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev* **14**(5): 559-573.

Nikiforov TT, Connolly BA. 1992. Oligodeoxynucleotides containing 4-thiothymidine and 6-thiodeoxyguanosine as affinity labels for the Eco RV restriction endonuclease and modification methylase. *Nucleic Acids Res* **20**(6): 1209-1214.

Noah JW, Lambowitz AM. 2003. Effects of maturase binding and Mg2+ concentration on group II intron RNA folding investigated by UV cross-linking. *Biochemistry* **42**(43): 12466-12480.

Noah JW, Park S, Whitt JT, Perutka J, Frey W, Lambowitz AM. 2006. Atomic force microscopy reveals DNA bending during group II intron ribonucleoprotein particle integration into double-stranded DNA. *Biochemistry* **45**(41): 12424-12435.

Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* **95**(19): 11163-11168.

Ornstein RL, Rein, R., Breen, D.L., Macelroy, R.D. 1978. An optimized potential function for the calculation of nucleic acid interaction energies: base stacking. *Biopolymers* **17**: 2341-2360.

Pabo CO, Sauer RT. 1992. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* **61**: 1053-1095.

Perron K, Goldschmidt-Clermont M, Rochaix JD. 1999. A factor related to pseudouridine synthases is required for chloroplast group II intron trans-splicing in Chlamydomonas reinhardtii. *EMBO J* **18**(22): 6481-6490.

Perutka J, Wang W, Goerlitz D, Lambowitz AM. 2004. Use of computer-designed group II introns to disrupt Escherichia coli DExH/D-box protein and DNA helicase genes. *J Mol Biol* **336**(2): 421-439.

Qin PZ, Pyle AM. 1998. The architectural organization and mechanistic function of group II intron structural elements. *Curr Opin Struct Biol* **8**(3): 301-308.

Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, Lambowitz AM. 1999. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry* **38**(28): 9069-9083.

San Filippo J. 2003. The DNA-binding and DNA endonuclease domains of a group II intron-encoded protein: characterization and application to the engineering of gene targeting vectors. In *College of Natural Sciences*, Vol Doctor of Philosphoy. University of Texas at Austin.

San Filippo J, Lambowitz AM. 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol* **324**(5): 933-951.

Sarafianos SG, Das K, Tantillo C, Clark AD, Jr., Ding J, Whitcomb JM, Boyer PL, Hughes SH, Arnold E. 2001. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *Embo J* **20**(6): 1449-1461.

Singh NN, Lambowitz AM. 2001. Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *J Mol Biol* **309**(2): 361-386.

Singh RN, Saldanha RJ, D'Souza LM, Lambowitz AM. 2002. Binding of a group II intron-encoded reverse transcriptase/maturase to its high affinity intron RNA binding site involves sequence-specific recognition and autoregulates translation. *J Mol Biol* **318**(2): 287-303.

Smith D, Zhong J, Matsuura M, Lambowitz AM, Belfort M. 2005. Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes Dev* **19**(20): 2477-2487.

Thingholm TE, Jorgensen TJ, Jensen ON, Larsen MR. 2006. Highly selective enrichment of phosphorylated peptides using titanium dioxide. *Nat Protoc* **1**(4): 1929-1935.

Till B, Schmitz-Linneweber C, Williams-Carrier R, Barkan A. 2001. CRS1 is a novel group II intron splicing factor that was derived from a domain of ancient origin. *RNA* **7**(9): 1227-1238.

Tocilj A, Schlunzen F, Janell D, Gluhmann M, Hansen HA, Harms J, Bashan A, Bartels H, Agmon I, Franceschi F et al. 1999. The small ribosomal subunit from Thermus thermophilus at 4.5 A resolution: pattern fittings and the identification of a functional site. *Proc Natl Acad Sci U S A* **96**(25): 14252-14257.

Treiber N, Treiber T, Zocher G, Grosschedl R. 2010. Structure of an Ebf1:DNA complex reveals unusual DNA recognition and structural homology with Rel proteins. *Genes Dev* **24**(20): 2270-2275.

Verschoor A, Warner JR, Srivastava S, Grassucci RA, Frank J. 1998. Three-dimensional structure of the yeast ribosome. *Nucleic Acids Res* **26**(2): 655-661.

Wang BH, Biemann K. 1994. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of chemically modified oligonucleotides. *Anal Chem* **66**(11): 1918-1924.

Wang MD, Yin H, Landick R, Gelles J, Block SM. 1997. Stretching DNA with optical tweezers. *Biophys J* **72**(3): 1335-1346.

Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell* **4**(2): 239-250.

Waugh RJ, Bowie JH, Gross ML. 1993. Collision-Induced Dissociations of Deprotonated Peptides - Dipeptides Containing Asn, Arg and Lys. *Aust J Chem* **46**(5): 693-702.

Waugh RJ, Bowie JH, Hayes RN. 1991. Collision-Induced Dissociations of Deprotonated Peptides - Dipeptides Containing Aspartic or Glutamic Acids. *Org Mass Spectrom* **26**(4): 250-256.

Waugh RJ, Eckersley M, Bowie JH, Hayes RN. 1990. Collision-Induced Dissociations of Deprotonated Peptides - Dipeptides Containing Serine or Threonine. *Int J Mass Spectrom* **98**(2): 135-145.

Weiner EM, Robson S, Marohn M, Clubb RT. 2010. The Sortase A enzyme that attaches proteins to the cell wall of Bacillus anthracis contains an unusual active site architecture. *J Biol Chem* **285**(30): 23433-23443.

Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**(10): 3353-3362.

Yao J, Lambowitz AM. 2007. Gene targeting in Gram-negative bacteria by use of a mobile group II intron ("Targetron") expressed from a broad-host-range vector. *Appl Environ Microbiol* **73**(8): 2735-2743.

Yao J, Zhong J, Fang Y, Geisinger E, Novick RP, Lambowitz AM. 2006. Use of targetrons to disrupt essential and nonessential genes in Staphylococcus aureus reveals temperature sensitivity of Ll.LtrB group II intron splicing. *RNA* **12**(7): 1271-1281.

Yao J, Zhong J, Lambowitz AM. 2005. Gene targeting using randomly inserted group II introns (targetrons) recovered from an Escherichia coli gene disruption library. *Nucleic Acids Res* **33**(10): 3351-3362.

Yousheng Hua SW, Yanan Yang, Lixin Shen, Yansan Xiong, Xiaoying Xu, Fagen Zhang, Judy L. Bolton, Richard B. van Breemen. 2000. Comparison of Negative

and Positive Ion Tandem Mass Spectrometry for the Liquid Chromatography Tandem Mass Spectrometry of Oxidized Deoxynucleosides. *J AM Soc for Mass Spect* **12**: 80-87.

Zhong J, Karberg M, Lambowitz AM. 2003. Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. *Nucleic Acids Res* **31**(6): 1656-1664.

Zhong J, Lambowitz AM. 2003. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *Embo J* **22**(17): 4555-4565.

Zhuang F, Mastroianni M, White TB, Lambowitz AM. 2009. Linear group II intron RNAs can retrohome in eukaryotes and may use nonhomologous end-joining for cDNA ligation. *Proc Natl Acad Sci U S A* **106**(43): 18189-18194.

Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM. 1995a. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. *Cell* **83**(4): 529-538.

Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995b. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82**(4): 545-554.

Zink SD, Burns DL. 2005. Importance of srtA and srtB for growth of Bacillus anthracis in macrophages. *Infect Immun* **73**(8): 5222-5228.